# CIUDADES INTELIGENTES Y DATOS ABIERTOS:
# UN DASHBOARD BASADO EN MINERÍA DE DATOS

**JOSE ARNULFO ESTEVEZ BLANCO**


**UNIVERSIDAD CATÓLICA DE COLOMBIA**
**FACULTAD DE INGENIERÍA**
**PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**
**TRABAJO DE GRADO**
**BOGOTÁ**
**2017**

# CIUDADES INTELIGENTES Y DATOS ABIERTOS: UN DASHBOARD BASADO EN MINERÍA DE DATOS

**JOSE ARNULFO ESTEVEZ BLANCO**

**Trabajo de grado como requisito para optar al título de Ingeniero de Sistemas y Computación**

**Director**
**Jhon Velandia MSc.**

**UNIVERSIDAD CATÓLICA DE COLOMBIA**
**FACULTAD DE INGENIERÍA**
**PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**
**TRABAJO DE GRADO**
**BOGOTÁ**
**2017**

creative commons

Atribución-NoComercial 2.5 Colombia (CC BY-NC 2.5)

La presente obra está bajo una licencia:
**Atribución-NoComercial 2.5 Colombia (CC BY-NC 2.5)**
Para leer el texto completo de la licencia, visita:
http://creativecommons.org/licenses/by-nc/2.5/co/

**Usted es libre de:**

Compartir - copiar, distribuir, ejecutar y comunicar públicamente la obra

hacer obras derivadas

**Bajo las condiciones siguientes:**

**Atribución** — Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciante (pero no de una manera que sugiera que tiene su apoyo o que apoyan el uso que hace de su obra).

**No Comercial** — No puede utilizar esta obra para fines comerciales.

Nota de aceptación

_____
_____
_____
_____
_____

_____
Presidente del Jurado

_____
Jurado

_____
Jurado

Bogotá, 21, noviembre, 2017

**CONTENIDO**

pág.

## LISTA DE FIGURAS

Pág.

pág.

## LISTA DE CUADROS

Pág.

## RESUMEN

La transparencia en la contratación pública ha venido tomando importancia en los últimos años debido a sucesos que se han venido presentando con varias contrataciones en el sector público, mediante el análisis de diversos conjuntos de datos de contratos suministrados por datos abiertos con un enfoque en SECOP se buscaron patrones que ayuden a determinar comportamientos por fuera de lo común que se presentan en la actualidad en Colombia. Se realizó un desarrollo progresivo de diferentes etapas que comprende desde la recopilación bibliográfica hasta el procesamiento de los datos extraídos desde datos abiertos, la elaboración de un modelo de minería de datos, análisis de los resultados y publicación en un dashboard. Para la realización de todas las etapas se utilizó la metodología CRISP-DM con el fin de obtener valor en los datos suministrados. Como resultado del análisis se encontró relación entre algunas regiones del país, el tipo de proceso de contratación, días de ejecución y el plazo de ejecución del contrato.



**Palabras clave**: Dashboard, Ciudades inteligentes, Contratos, Datos abiertos, Minería de datos, Agrupación / Clustering.

# INTRODUCCIÓN

En la actualidad existe dispersión de la información en las fuentes de las ciudades generando un estado incompleto, limitando las mejoras en sectores críticos por el desconocimiento, por ejemplo, en el ámbito de la contratación pública sin un estado claro no es posible identificar los problemas de corrupción, ayudaría poder reconocer patrones en los contratos contribuyendo a la mitigación de estas acciones.

La minería de datos se utiliza en diferentes unidades policiales e inspecciones especiales, cuya misión es identificar actividades fraudulentas y descubrir tendencias delictivas. Por ejemplo, “estas metodologías pueden ayudar a los analistas en la identificación de patrones críticos de conducta, en las interacciones de comunicación de las organizaciones de narcóticos, en las transacciones monetarias de lavado de dinero y operaciones de información privilegiada”[1].

Los Dashboards en las ciudades miden el progreso de las ciudades utilizando métricas urbanas mejorando sus estrategias de gestión a medida que se disponga de nuevos datos, mediante la toma de decisiones basadas en datos. Por ejemplo, lo realizado en ciudades como “Chicago, new York, Londres entre otras ciudades de Estados unidos y Europa donde se evidencia mejora en la transparencia y la responsabilidad con la ciudad”[2].

[1] GOMEZ ZOTANO, Miguel Angel y BERSINI, Hugues. A Data-driven Approach to Assess the Potential of Smart Cities: The Case of Open Data for Brussels Capital Region. En: Energy Procedia. Marzo, 2017. vol. 111, p. 750-758.

[2] MARTIN, Erika G y BEGANY, Grace M. Opening government health data to the public: benefits, challenges, and lessons learned from early innovators. En: Journal of the American Medical Informatics Association. Agosto – septiembre, 2016. vol. 24, no. 2, p. 345-351.

# 1. PLANTEAMIENTO DEL PROBLEMA

## 1.1 DESCRIPCIÓN DEL PROBLEMA

“Los datos e información pública masiva generan la posibilidad de mejorar la eficiencia a la hora de identificar, analizar y proponer soluciones a problemas urbanos”[3].

Al comprender las técnicas de minería de datos para obtener información útil se da lugar a la construcción de un dashboard con la posibilidad que los gobernantes, pensadores públicos, “miembros de movimientos ciudadanos o ONG de la región puedan tomar, influir o informar sus decisiones basándose en los datos que dispone”[4].

Los datos abiertos aportan a “la tendencia mundial de toma de decisiones con base en observaciones empíricas y análisis cuantitativos con el fin de poder ayudar a resolver los problemas urbanos más latentes e importantes de la región”[5].

## 1.2 FORMULACIÓN DEL PROBLEMA

El objetivo primordial es predecir el comportamiento de variables partiendo desde datos actuales proporcionados por la fuente de datos abiertos datos.gov.co, permitiendo la toma de decisiones con la información producto del modelo o método utilizado de minería de datos.

Los resultados del proceso se incluirán en un dashboard que permite ver el estado futuro de la ciudad con base en datos actuales, llevando a la interrogante ¿Qué método de minera de datos aplicar a los datos para identificar comportamientos en la información suministrada de contratación pública que ayude en el mejoramiento de la transparencia?

[3] GOMEZ ZOTANO y BERSINI, Op. cit., p. 750
[4] MARTIN y BEGANY, Op. cit., p. 345
[5] GOMEZ ZOTANO y BERSINI, Op. cit., p. 751

# 2. OBJETIVOS

## 2.1 OBJETIVO GENERAL

Implementar un dashboard basado en ciudades inteligentes en la ciudad de Bogotá utilizando datos abiertos relacionados con procesos contractuales.

## 2.2 OBJETIVOS ESPECÍFICOS

- Identificar los métodos de minería de datos para encontrar información desconocida que apoye la optimización de procesos contractuales.

- Diseñar una arquitectura de software que satisfaga los requisitos funcionales y no funcionales del dashboard.

- Desarrollar un prototipo de software considerando la arquitectura propuesta.

- Evaluar la calidad de los datos suministrados por el prototipo.

## 3. ALCANCES Y LIMITACIONES

- Se utilizará solo los datos proporcionados por www.datos.gov.co siguiendo los lineamientos del macro proyecto.

- El proyecto abarcara datos asociados a contratos (SECOP) siguiendo los lineamientos del macro proyecto.

- El proyecto se desarrollará en 16 semanas correspondientes al periodo académico 2017-3.

- Los datos que van a ser utilizados serán guardados por medio de otra herramienta a la base de datos mongo-db desarrollada siguiendo los lineamientos del macro proyecto.

- El objetivo inicial de predecir variables no es posible cumplirlo a causa de que no se tienen los datos necesarios como lo es un histórico amplio, solo se tienen de los años 2015, 2016 y fracción del 2017. El nuevo objetivó será el descubrimiento de datos y análisis descriptivo de los mismos.

- Debido a limitaciones de la herramienta de extracción solo se guardan 50.000 registros siendo esta la cantidad inicial de datos con los que se van a realizar el análisis.

- El conjunto de datos analizados será a nivel Colombia para obtener un contexto amplio y no solo del 15% pertenecen a contratos en Bogotá D.C.

- Los conjuntos de datos con un alto número de registros no es posible analizarlo con Rapidminer debido a limitaciones de hardware y licenciamiento que solo permite realizar procesos con hasta 10.000 registros y 2 GB de memoria en su versión gratuita.

- El uso de la base de datos mongo-db con Tableau restringe algunas funcionalidades como por ejemplo el manejo de fechas correctamente, por lo tanto, se debe guardar un archivo con los datos desde la aplicación para acceder a todas las funcionalidades que ofrece.

# 4. METODOLOGÍA

## 4.1 METODOLOGÍA CRISP-DM

Un grupo de organizaciones involucradas en la minería de datos propusieron una guía de referencia para desarrollar proyectos de minería de datos, denominados "CRISP-DM (Cross Industry Standard Process for Data Mining). CRISP-DM es considerado el estándar para desarrollar proyectos de minería de datos y descubrimiento de conocimiento"[6].

> La metodología de minería de datos CRISP-DM se describe en términos de un modelo de proceso jerárquico, que consiste en conjuntos de tareas descritas en cuatro niveles de abstracción de lo general a lo específico. En el nivel superior, el proceso de minería de datos se organiza en una serie de fases; Cada fase consta de varias tareas genéricas de segundo nivel. Este segundo nivel se llama genérico, porque se pretende que sea lo suficientemente general para cubrir todas las posibles situaciones de minería de datos. El tercer nivel, el nivel de tarea especializado, es el lugar para describir cómo las acciones en las tareas genéricas deben llevarse a cabo en ciertas situaciones específicas. El cuarto nivel, la instancia de proceso, es un registro de las acciones, decisiones y resultados de un compromiso de minería de datos real[7].

Horizontalmente, la metodología CRISP-DM distingue entre el modelo de referencia y la guía del usuario. El modelo de referencia presenta una visión general rápida de las fases, las tareas y sus resultados y describe qué hacer en un proyecto de minería de datos. "La guía del usuario proporciona consejos y sugerencias más detalladas para cada fase y cada tarea dentro de una fase y describe cómo realizar un proyecto de minería de datos"[8]. CRISP-DM distingue entre cuatro dimensiones diferentes de los contextos de minería de datos:

- El dominio de la aplicación es el área específica en la que tiene lugar el proyecto de minería de datos.

- El tipo de problema de minería de datos describe las clases específicas de objetivos con los que se trata el proyecto de minería de datos.

- El aspecto técnico cubre problemas específicos en la minería de datos que describen los diferentes desafíos técnicos que normalmente ocurren durante la minería de datos.

---

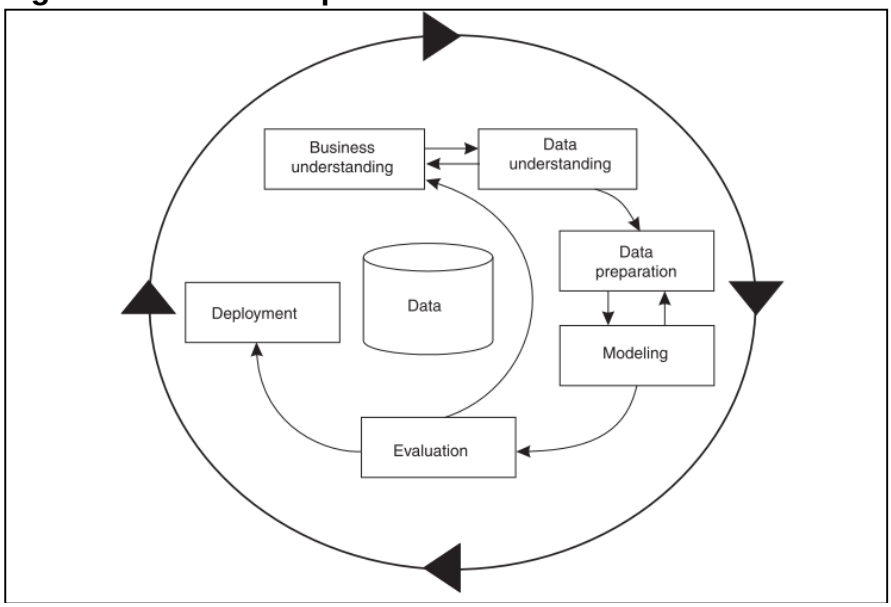
[6] MARISCAL, Gonzalo; MARBÁN, Óscar y FERNÁNDEZ, Covadonga. A survey of data mining and knowledge discovery process models and methodologies. En: The Knowledge Engineering Review. Junio – agosto, 2010. vol. 25, no. 2, p. 137.
[7] Ibíd., p. 138
[8] Ibíd., p. 138

➢La dimensión especifica de la herramienta y la técnica, cuya herramienta o técnica se aplican durante el proyecto de minería de datos[9].

El modelo de proceso CRISP-DM para la minería de datos proporciona una visión general del ciclo de vida de un proyecto de minería de datos. Contiene las fases correspondientes de un proyecto, sus tareas respectivas y las relaciones entre estas tareas. “El ciclo de vida de un proyecto de minería de datos de acuerdo con CRISP-DM consta de seis fases”[10] (véase la Figura 3).

La secuencia de las fases no es estricta. Siempre se requiere ir y venir entre diferentes fases. Depende del resultado de cada fase, qué fase o qué tarea particular de una fase debe realizarse a continuación. Las flechas indican las dependencias más importantes y frecuentes entre las fases.

**Figura 1. Modelo del proceso CRISP-DM**




Fuente. CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ, Thomas; SHEARER, Colin y WIRTH, Rüdiger. CRISP-DM 1.0. Step-by-step data mining guide. Pittsburgh: The CRISP-DM consortium, 2000. p. 10

A continuación, se realiza una corta descripción de las fases:

➢**Comprensión del negocio**. Esta fase inicial se centra en la comprensión de los objetivos del proyecto y los requisitos desde una perspectiva de negocio, y luego convertir este conocimiento en una definición de problema de minería de datos, y un plan preliminar diseñado para lograr los objetivos.

[9] Ibíd., p. 139
[10] Ibíd., p. 139

➢**Comprensión de los datos.** La fase de comprensión de los datos comienza con una recopilación de datos inicial y prosigue con las actividades para familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir las primeras percepciones de los datos o detectar subconjuntos interesantes para formar hipótesis de información oculta.

➢**Preparación de los datos**. La fase de preparación de los datos abarca todas las actividades para construir el conjunto final de datos (datos que se introducirán en la(s) herramienta(s) de modelado) a partir de los datos iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en un orden prescrito. Las tareas incluyen selección de tabla, registro y atributo, así como transformación y limpieza de datos para herramientas de modelado.

➢**Modelado**. En esta fase, se seleccionan y aplican diversas técnicas de modelado, y sus parámetros se calibran a valores óptimos. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, a menudo se necesita regresar a la fase de preparación de datos.

➢**Evaluación.** En esta etapa del proyecto se ha construido un modelo o modelos que parece tener alta calidad, desde una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluar más a fondo el modelo y revisar los pasos ejecutados para construir el modelo, para asegurarse de que alcanza adecuadamente los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión comercial importante que no se ha considerado suficientemente. Al final de esta fase, se debe llegar a una decisión sobre el uso de los resultados de la extracción de datos.

➢**Despliegue.** Generalmente, la creación del modelo no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, será necesario organizar el conocimiento extraído, así como presentarlo de manera útil al cliente. Dependiendo de los requisitos, la fase de despliegue puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible. En muchos casos, será el cliente, no el analista de datos, quien llevará a cabo las etapas de despliegue. Sin embargo, incluso si el analista no llevará a cabo el esfuerzo de despliegue, es importante que el cliente entienda por adelantado qué acciones deberán llevarse a cabo con el fin de hacer uso de los modelos creados[11].

[11] Ibíd., p. 140

# 5. MARCO REFERENCIAL

## 5.1 MARCO TEÓRICO

**5.1.1 Datos Abiertos**. En un sentido amplio, el sector público es la fuente de una enorme cantidad de datos creado o recopilados como parte de sus funciones, por ejemplo, “horarios de transporte público, gobierno estadísticas, catálogos de bibliotecas o museos, mapas, información sobre ingresos y gastos, concursos públicos, entre otros datos relevantes”[12].

Los datos abiertos son todos los datos accesibles y reutilizables, sin exigencia de permisos específicos consolidados en una sola fuente de información con una amplia gama de información útil para mejorar la vida en la ciudad.

**5.1.2 DashBoard**. Los dashboard son componentes comunes de la mayoría de los sistemas de gestión del rendimiento, los sistemas de medición del rendimiento, las suites de BPM y las plataformas de BI. “Los dashboard proporcionan visualizaciones visuales de información importante que se consolida y se organiza en una sola pantalla para que la información pueda ser digerida de una sola mirada y fácilmente explorada”[13].

**5.1.3 Ciudades Inteligentes**. El éxito de los proyectos de ciudad inteligente está intrínsecamente relacionado con la existencia de grandes volúmenes de datos que podrían ser procesados para alcanzar sus objetivos, ciudades o países como “Vancouve, Portland (Oregon), San Francisco (2009), Nueva York (2012) o Nueva Zelanda (2011) tienen sitios webs donde publican datos sin requieren licencia facilitando su acceso y explotación”[14].

La economía mundial se ha centrado en los datos y, en consecuencia, aquellos con capacidades para extraer el máximo beneficio de sus datos tendrá el poder en política, social, cultural y, especialmente, el nivel económico. Por lo tanto, en los últimos años, un número creciente de gobiernos ha comenzado a abrir sus datos. “Este movimiento llamado gobierno abierto ha resultado en el lanzamiento de numerosos portales de datos abiertos y de infraestructuras que tienen como objetivo proporcionar un punto único de acceso a datos del gobierno y explorar sus consecuencias”[15].

---

[12] MARTIN y BEGANY, Op. cit., p. 346

[13] RUDY; MIRANDA, E., y SURYANI, E. Implementation of datawarehouse, datamining and dashboard for higher education. En: Journal of Theoretical and Applied Information Technology. Junio – julio, 2014. vol. 64, no. 3., p. 710

[14] GOMEZ ZOTANO y BERSINI, Op. cit., p. 752

[15] MÁCHOVÁ, R., y LNENICKA, M. Evaluating the Quality of Open Data Portals on the National Level. En: Journal of Theoretical and Applied Electronic Commerce Research. Mayo – junio, 2017. vol. 12, no. 1, p. 21.

Las ciudades inteligentes buscan optimizar sus procesos con el objetivo de "mejorar la calidad de vida de sus habitantes, identificando los posibles problemas de los que se aquejan y brindándoles solución prontamente"[16].

**5.1.4 Minería de Datos**. En general, en el mundo industrial y de prensa, la minería de datos se utiliza para referirse a todo el proceso KDD. Por lo tanto, ambos términos se pueden utilizar indistintamente cuando se refieren a esta área. Últimamente, "el término minería de datos y descubrimiento de conocimiento ha sido propuesto como el nombre más adecuado para el proceso general de KDD"[17].

A principios de los años 90, cuando se creó el termino KDD, se desarrollaron algoritmos de minería de datos capaz de resolver todos los problemas relacionados con la búsqueda de conocimientos útiles en grandes volúmenes de datos. Aparte de desarrollar algoritmos, algunas herramientas específicas, tales como: "Clementine, Weka, IBM Intelligent Miner también se desarrollaron para simplificar la aplicación de algoritmos de minería de datos y proporcionar algún tipo de apoyo para todas las actividades involucradas en el Procesos de Extracción de Conocimiento-KDD"[18].

Algunos métodos utilizados son: sequential pattern mining, clustering-based classification trees, hybrid learning y flexible pattern mining.

En minería de datos encontramos dos tareas principales que son descripción y predicción:

- **Descripción.** Los métodos descriptivos buscan patrones interpretables para describir datos. Son los siguientes: clustering, descubrimiento de reglas de asociación y descubrimiento de patrones secuenciales. Los métodos descriptivos se han utilizado, por ejemplo, para ver qué productos se adquieren conjuntamente en el supermercado.

- **Predicción.** En la predicción el modelo debe inferir una variable a partir de alguna combinación de otras variables incluidas en los datos. La predicción requiere etiquetas para la variable de salida para un conjunto de datos limitado, donde una etiqueta suponga una información fiable sobre el valor de la variable de salida en casos específicos, en algunos casos es importante considerar el grado en el que estas etiquetas puedan ser aproximadas o inciertas. En algunos casos, métodos de predicción pueden ser usados para estudiar qué características de un modelo son importantes para una predicción, dando información sobre la construcción subyacente. Este es un enfoque común en programas de investigación que tratan

[16] Ibíd., p. 22
[17] MARISCAL; MARBÁN y FERNÁNDEZ, Op. cit., p. 141
[18] RIQUELME, J.C.; RUIZ, R. y GILBERT, K. Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial, En: Revista Iberoamericana de Inteligencia Artificial. Febrero – marzo, 2006. no, 29, p. 12

de predecir resultados sin predecir anteriormente factores intermedios. En un segundo tipo de uso, los métodos de predicción son utilizados para predecir cuál será el valor de salida en contextos donde no es deseable obtener una etiqueta para esa construcción (por ejemplo, en ocasiones en las que no haya datos etiquetados) [19].

**5.1.4.1 Procesos de Extracción de Conocimiento.** El proceso de Extracción de conocimiento-KDD involucra numerosos pasos con muchas decisiones tomadas por el usuario. Brachman y Anand (1996) dan una visión práctica del proceso KDD, enfatizando la naturaleza interactiva del proceso. Sus pasos básicos son los siguientes (véase la Figura 2).

**Figura 2. Pasos KDD**


Fuente. El Autor

Primero está desarrollar una comprensión del dominio de aplicación y el conocimiento previo relevante e identificar la meta del proceso de KDD desde el punto de vista del cliente.

Segundo, crear un conjunto de datos de destino: seleccionar un conjunto de datos o centrarse en un subconjunto de variables o muestras de datos en las que se debe

[19] JIMÉNEZ GALINDO, Álvaro y ÁLVAREZ GARCÍA, Hugo. Minería de Datos en la Educación [en línea]. Madrid: Revista Inteligencia En Redes de Comunicación [citado 15 agosto, 2017]. Disponible en Internet: <URL: https://www.it.uc3m.es/jvillena/irc/practicas/10-11/08mem.pdf>

realizar el descubrimiento.

La tercera es la limpieza de datos y el pre-procesamiento. Las operaciones básicas incluyen eliminar el ruido, recopilar la información necesaria para modelar o contabilizar el ruido, decidir estrategias para manejar los campos de datos faltantes y contabilizar la información de la secuencia temporal y los cambios conocidos.

Cuarto es la reducción y proyección de datos: encontrar características útiles para representar los datos dependiendo del objetivo de la tarea. Con los métodos de reducción de la dimensionalidad o de transformación, se puede reducir el número efectivo de variables bajo consideración, o se pueden encontrar representaciones invariantes para los datos.

En quinto lugar, corresponden a los objetivos del proceso KDD del paso 1 con un método particular de minería de datos. Por ejemplo, clasificación, regresión, agrupamiento, modelos de dependencias, detección de cambios y desviaciones, asociación y análisis de evolución entre otros.

El sexto es el análisis exploratorio y la selección de modelos e hipótesis: la elección del algoritmo de minería de datos y el método o métodos de selección que se utilizarán para buscar patrones de datos. Este proceso incluye decidir qué modelos y parámetros pueden ser apropiados y hacer coincidir un método particular de minería de datos con los criterios generales del proceso KDD.

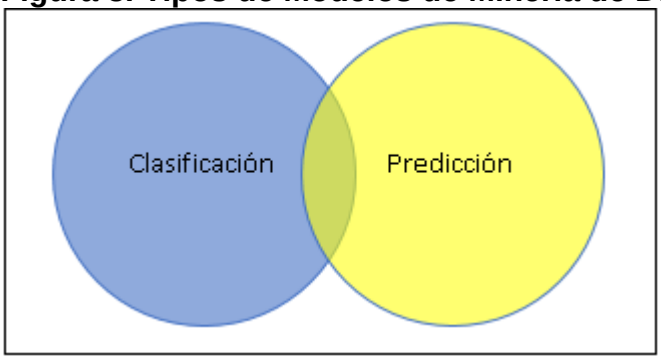
La séptima es la minería de datos: buscar patrones de interés en una forma representativa particular o un conjunto de tales representaciones, incluyendo reglas de clasificación o árboles, regresión y agrupación.

La octava es la interpretación de patrones minados, posiblemente regresando a cualquiera de los pasos 1 a 7 para iteración adicional. Este paso también puede implicar la visualización de los patrones y modelos extraídos o la visualización de los datos extraídos por el modelo[20].

Noveno actuar sobre el conocimiento descubierto: “usando el conocimiento directamente, incorporando el conocimiento en otro sistema para acciones futuras, o simplemente documentándolo y reportándolo a las partes interesadas. Este proceso también incluye buscar y resolver conflictos potenciales con conocimiento previamente creado (o extraído)”[21] (véase la Figura3).

[20] MÁCHOVÁ y LNENICKA, Op. cit., p. 22.
[21] FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). From Data Mining to Knowledge Discovery in Databases. En: AI Magazine. Junio – julio, 1996. vol. 17, no. 3, p. 37.

**Figura 3. Tipos de Modelos de Minería de Datos Principales**




Fuente. El Autor

Los métodos o modelos aplicados en minería de datos se dividen en dos principales, métodos predictivos y métodos descriptivos. En el extremo predictivo, el objetivo de la minería de datos es producir un modelo, que se puede utilizar para realizar la clasificación, predicción, estimación u otras tareas similares. En el extremo descriptivo del espectro, el objetivo es obtener una comprensión del sistema analizado mediante el descubrimiento de patrones y relaciones en grandes conjuntos de datos. "La importancia relativa de la predicción y la descripción para aplicaciones específicas de minería de datos puede variar considerablemente. Las metas de predicción y descripción se consiguen mediante técnicas de minería de datos"[22].

Cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido, junto con los datos históricos de esos ejemplos. " Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos. Cuando este modelo se aplica a nuevas entradas de datos, el resultado es una predicción del comportamiento futuro de los mismos"[23].

La mayoría de las técnicas de Minería de Datos son adecuadas para usar la predicción a partir de datos históricos como también de datos de entrenamiento de forma adecuada. "La elección de la técnica depende de la naturaleza de los datos de entrada, el tipo de valor que se predice, y la importancia concedida a la explicación de la predicción"[24].

---

[22] FU, Tak-chung. A review on time series data mining.en: Engineering Applications of Artificial Intelligence. February, 2011. vol. 24, no. 1, p. 164

[23] VALENCIA ZAPATA, Gustavo Adolfo. Minería de datos la minería de datos como herramienta para la toma de decisiones estratégicas. Bogotá: IMG, 2010. p. 8

[24] Ibíd., p. 8

Dado que la Minería de Datos es un campo interdisciplinar, existe un conjunto de tareas que cumplen con sus propósitos y que pueden ser utilizadas en áreas de aplicación específicos, “diversas técnicas de Minería de Datos se utilizan para llevar a cabo las tareas de la misma, estas técnicas consisten en algoritmos específicos que pueden ser utilizados para cada función”[25]

**5.1.5 Algoritmos de minería de datos.** En minería de datos existen una variedad de algoritmos que se adaptan a diferentes finalidades, a continuación, se listan los de mayor relevancia:

**5.1.5.1 Rule Induction**. La inducción de reglas es un proceso de minería de datos que consiste en deducir reglas if-then de un conjunto de datos. Estas reglas de decisión simbólica explican una relación inherente entre los atributos y las etiquetas de clase en el conjunto de datos. Muchas experiencias de la vida real se basan en la inducción de reglas intuitivas.

La inducción de la regla proporciona un poderoso enfoque de clasificación que puede ser fácilmente entendido por la audiencia general. Aparte de su uso en el Análisis Predictivo mediante la clasificación de datos desconocidos, la inducción de reglas también se utiliza para “describir los patrones en los datos. La descripción está en forma de simples reglas if-then que pueden ser fácilmente entendidas por los usuarios comunes”[26].

**5.1.5.2 k-Nearest Neighbors**. La minería predictiva de datos usando árboles de decisión y técnicas de inducción de reglas se construyeron generalizando la relación dentro del conjunto de datos y usándolo para predecir el resultado de nuevos datos no vistos.

> Estos enfoques se llaman aprendices ansiosos porque intentan encontrar una mejor aproximación de la relación real entre las variables de entrada y de destino. Pero también hay un enfoque alternativo simple.
>
> Esta clase de alternativa de estudiantes adopta un enfoque contundente, donde no se realiza "aprendizaje" a partir del conjunto de datos de entrenamiento; Más bien el conjunto de datos de entrenamiento se utiliza como una tabla de búsqueda para coincidir con las variables de entrada y encontrar el resultado, a esto se le llama k-vecinos más cercanos[27].

[25] Ibíd., p. 10.
[26] KOTU, Vijay y DESHPANDE, Bala. Predictive analytics and data mining: concepts and practice with RapidMiner. Massachusetts, United States of America: Elliot Steven, 2015. p. 13. ISBN: 978-0-12-801460-8.
[27] Ibíd., p. 99.

**5.1.5.3 Naïve Bayesian**. El algoritmo de Bayes encuentra sus raíces en la estadística y la teoría de probabilidades.

> En general, las técnicas de clasificación tratan de predecir las etiquetas de clase basadas en atributos mejor aproximando la relación entre variables de entrada y salida. Cada día, estimamos mentalmente una miríada de resultados basados en pruebas pasadas.
>
> Los clasificadores bayesianos también han mostrado alta precisión y rapidez cuando se aplican a bases de datos grandes. Los clasificadores bayesianos Naïve suponen que el efecto de un valor de atributo en una clase dada es independiente de los valores de los otros atributos. Esta suposición se llama independencia condicional de clase[28].

➢**Artificial Neural Networks**. El objetivo de un algoritmo de análisis predictivo es modelar la relación entre variables de entrada y salida. La técnica de red neural se acerca a este problema desarrollando una explicación matemática que se asemeja mucho al proceso biológico de una neurona. Aunque los desarrolladores de esta técnica han utilizado muchos términos biológicos para explicar el funcionamiento interno del proceso de modelado de redes neuronales, tiene una base matemática simple. Consideremos el modelo matemático lineal simple:

$$y = 1 + 2X_1 + 3X_2 + 4X_3$$

➢**Support Vector Machines**. Las Máquinas de Soporte Vectorial, es un método para la clasificación de datos lineales y no lineales. En pocas palabras, una máquina de vector de soporte (o SVM) es un algoritmo que funciona de la siguiente manera. “Utiliza un mapeo no lineal para transformar los datos de entrenamiento originales en una dimensión superior. Dentro de esta nueva dimensión, busca el hiperplano de separación óptimo lineal (es decir, un "límite de decisión" que separa las tuplas de una clase de otra)”[29].

**5.1.5.4 Ensemble Learners**. Métodos de conjunto o aprendices optimizar el problema de la búsqueda de hipótesis mediante el empleo de una matriz de modelos de predicción individuales y luego combinarlos para formar una hipótesis o modelo agregado. “Estos métodos proporcionan una técnica para generar una mejor hipótesis combinando múltiples hipótesis en una sola”[30].

---

[28] Ibíd., p. 111.
[29] HAN, Jiawei y KAMBER, Micheline. Data Mining: Concepts and Techniques. United States of America: Morgan Kaufmann, 2006. p. 337.
[30] KOTU y DESHPANDE. Op. cit., p. 148.

Debido a que una sola hipótesis puede ser localmente óptima o sobrevalorar un conjunto de entrenamiento particular, “la combinación de varios modelos puede mejorar la precisión forzando una solución de meta-hipótesis. Se puede demostrar que en ciertas condiciones este poder predictivo combinado es mejor que el poder predictivo de los modelos individuales”[31].

**5.1.5.5 Linear Regression**. El análisis de regresión lineal implica una variable de respuesta, y, y una variable predictora única, x. Es la forma más simple de regresión, y los modelos y como una función lineal de x. Es decir,

$$y = b + wx$$

Donde la variación de y se supone que es constante, y bw son coeficientes de regresión que especifican la intersección en Y y la pendiente de la línea, respectivamente. “Los coeficientes de regresión, w y b, también pueden considerarse como pesos”[32].

**5.1.5.6 Logistic Regression**. La regresión logística modela la probabilidad de que algún evento ocurra como una función lineal de un conjunto de variables predictoras. “Los datos de recuento frecuentemente presentan una distribución de Poisson y son comúnmente modelados usando la regresión de Poisson”[33].

**5.1.5.7 Apriori**. El algoritmo Apriori aprovecha algunos principios lógicos simples en los conjuntos de elementos de celosía para reducir el número de conjuntos de elementos a ensayar para la medida de soporte. “Los principios Apriori establecen que "Si un conjunto de elementos es frecuente, entonces todos sus subconjuntos serán frecuentes.". El conjunto de elementos es "frecuente" si el soporte para el conjunto de elementos es más que el umbral de soporte”[34].

**5.1.5.8 FP-Growth**. Se llama crecimiento de patrón frecuente, o simplemente crecimiento de PF, que adopta una estrategia de división y conquista como sigue. En primer lugar, comprime la base de datos que representa los elementos frecuentes en un árbol de patrón frecuente, o árbol FP, que retiene la información de asociación de conjuntos de elementos. “A continuación, divide la base de datos comprimida en un conjunto de bases de datos condicionales (un tipo especial de base de datos proyectada), cada una asociada con un elemento frecuente o "fragmento de patrón", y extrae cada base de datos por separado”[35].

---

[31] KOTU y DESHPANDE. Op. cit., p. 148.
[32] HAN y KAMBER. Op. cit., p. 81.
[33] Ibíd., p. 358.
[34] KOTU y DESHPANDE. Op. cit., p. 202.
[35] HAN y KAMBER. Op. cit., p. 245.

**5.1.5.9 k-Means**. El algoritmo K-means es muy sensible en términos de selección de los medios iniciales. “El método K-means se aplica cuando se define la media de un conjunto de objetos. La desventaja de K-means es que, no hay una respuesta específica para encontrar el número mínimo de clusters para cualquier conjunto de datos dado. Una solución, ya que es comparar los resultados de múltiples ejecuciones con diferentes clusters y elegir la mejor según criterios. Para la predicción de incendios forestales es utilizado el algoritmo de K-means”[36].

**5.1.5.10 DBSCAN**. “DBSCAN (Density-based Spatial Clustering de aplicación con ruido) se propuso que la conectividad de densidad de acceso para el manejo de la forma aleatoria de clúster y el ruido”[37].

DBSCAN es un algoritmo de agrupación basado en densidad. “El algoritmo aumenta regiones con densidad suficientemente alta en grupos y descubre racimos de forma arbitraria en bases de datos espaciales con ruido. Define un clúster como un conjunto máximo de puntos conectados a la densidad”[38].

**5.1.5.11 Self-Organizing Maps**. Un mapa de auto organización (SOM) es una poderosa técnica de agrupamiento visual que evolucionó a partir de la combinación de redes neuronales y clustering basado en prototipos. Una SOM es una forma de red neuronal donde la salida es una matriz visual organizada, usualmente una cuadrícula bidimensional con filas y columnas. “El objetivo de esta red neuronal es transferir todos los objetos de datos de entrada con n atributos (n dimensiones) a la red de salida de tal manera que los objetos uno junto al otro esté estrechamente relacionados entre sí”[39].

## 5.2 MARCO CONCEPTUAL

Desde principios del siglo XX, los gobiernos han utilizado indicadores sociales y económicos, como la tasa de desempleo, el producto interno bruto (PIB), el producto nacional bruto (PNB), la balanza de pagos, la inflación y el índice de precios al consumidor (IPC) para evaluar cómo se está desempeñando una nación.

> Del mismo modo, en la era posterior a la Segunda Guerra Mundial, muchas agencias como la Organización Mundial de la Salud (OMS), la Organización para la Cooperación Económica Operación y Desarrollo (OCDE) y el Programa de las Naciones Unidas para el Desarrollo (PNUD) miden, cotejan y rastrean el desempeño y la productividad de diversos fenómenos de salud, económicos y sociales en las

---

[36] SHAH, Chintan; y JIVANI, Anjali. Comparison of data mining clustering algorithms En: Nirma: University International Conference on Engineering (NUiCONE) Ahmedabad, India: IEEE, 2013. p. 1.
[37] Ibíd., p. 2.
[38] HAN y KAMBER. Op. cit., p. 218.
[39] KOTU y DESHPANDE. Op. cit., p. 242.

> naciones y regiones. En las últimas dos décadas, el uso de indicadores ha aumentado en todos los sistemas del sector público, incluida la administración pública y los servicios públicos, como la salud y la educación, y se utiliza para supervisar y evaluar diversos aspectos de las ciudades, como la competitividad, la sostenibilidad, calidad de vida, bienestar y servicios urbanos. Muchas ciudades de todo el mundo generan rutinariamente conjuntos de indicadores, utilizándolos para trazar el desempeño, guiar la formulación de políticas e informar cómo se gobiernan y regulan las ciudades[40]

En la ciudad de Bandung City una de las grandes ciudades de Indonesia se propuso una la aplicación para el monitoreo de la ciudad en un tablero único para ayudar a resumir la condición actual de la ciudad. El sistema de arquitectura utiliza sensores de red que consiste en nodos con sensores que tienen la función de capturar las condiciones de la ciudad, como la temperatura, la contaminación del aire, la contaminación del agua y la situación del tráfico. También "permite agregar otra información sobre la situación socioeconómica como servicio de salud pública, indicador económico y suministros de energía. Se realizó con éxito el desarrollado del prototipo del tablero de instrumentos de la ciudad inteligente para proporcionar información más"[41].

Los paneles dinámicos a menudo se exhiben en monitores de computadoras en salas de control modernas, resumiendo gráficamente un sistema en flujo para operadores humanos, o se encuentran a veces en espacios públicos para comunicar información a los ciudadanos, por ejemplo, el iPad en la oficina del alcalde de Londres. "Desde las emisiones de aterrizaje de la Luna de la NASA, tales pantallas se han convertido en representaciones icónicas y sublimes de cómo se mantiene el control"[42].

> Las salas de control urbanas de hoy generalmente se relacionan con un único dominio, como el tráfico o la seguridad o el clima. Sin embargo, en el caso de Río de Janeiro, los flujos de datos en vivo provienen de 30 agencias, incluidos tráfico y transporte público, servicios municipales y de servicios públicos, servicios de emergencia, información meteorológica e información enviada por los empleados y el público por teléfono, internet y radio, se alimentan en un único centro de análisis de datos donde se visualizan en una variedad de formas. En los últimos años se han estado desarrollando centros similares en otros lugares, acompañados de una gama de aplicaciones para que los ciudadanos accedan y utilicen algunos flujos de datos. Con respecto a este último, se han desarrollado varios prototipos de cuadros de

---

[40] KITCHIN, R.; LAURIAULT, T.P. y MCARDLE, G. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. En: Regional Studies, Regional Science. Junio – agosto, 2015. vol. 2, no. 1, p. 6

[41] SUAKANTO, S.; SUPANGKAT, S.H.; SUHARDI, A. y SARAGIH, R. Smart city dashboard for integrating various data of sensor networks En. Tangerang: International Conference on ICT for Smart Society (2, 13-14 Junio: Jakarta, Indonesia). Jakarta: IEEE, 2013. p. 1-2.

[42] KITCHIN; LAURIAULT y MCARDLE, Op. cit., p. 7

mando de acceso abierto en tiempo real, algunos de que combina visualizaciones de indicadores tradicionales con nuevos datos en tiempo real para proporcionar una visión global de la ciudad compuesta de datos administrativos y operativos, por ejemplo, Dublin Dashboard, un panel analítico, que se lanzó en septiembre de 2014. El poder de estos tableros radica en que brindan de manera rápida y efectiva a los administradores de las ciudades y, en menor medida, a los ciudadanos información actualizada sobre los diferentes aspectos de los sistemas y entornos urbanos, y cómo están cambiando en el tiempo y el espacio[43].

**5.2.1 Exploración de los Datos**. La mayoría de los algoritmos de minería de datos requieren que los datos se estructuren en un formato tabular con registros en filas y atributos en columnas.

La preparación de los datos comienza con una exploración en profundidad de los datos y una mayor comprensión del conjunto de datos. En general, un conjunto de datos originado para responder a una pregunta de negocio tiene que ser analizado, preparado y transformado antes de aplicar algoritmos y crear modelos.

Es importante verificar los datos usando técnicas de exploración de datos además del conocimiento previo de los datos y del negocio antes de la construcción de modelos para garantizar un cierto grado de calidad de los datos.

Algunos análisis descriptivos: Cálculo de mínimos y máximos, calcular media y la desviación estándar y examinar la distribución de los datos.

Una gráfica línea, grafica de barras y gráficos de dispersión son parte de las técnicas de exploración de datos que se utilizan en el entorno empresarial cotidiano, revelando mucha información sobre los datos, que puede utilizarse para decidir sobre los próximos pasos para extraer los datos[44].

**5.2.2 Técnica Arboles de decisión**. Esta técnica se aplica a la clasificación y predicción. Los árboles de decisión son ampliamente usados y pueden ser fácilmente explicados basándose en el criterio usado para dividir los datos en las extremidades del árbol.

Los árboles de decisión son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos.

Las técnicas basadas en árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos, los valores faltantes y los datos incongruentes que se puedan presentar en el conjunto de datos. Son bastante eficientes y obtienen resultados para clasificación, los métodos obtenidos se pueden expresar como conjuntos de reglas. Uno de los inconvenientes de los

[43] Ibíd., p. 7
[44] KOTU y DESHPANDE. Op. cit., p. 23-27.

árboles de decisión es su limitada expresividad y que son inestables ante variaciones de la muestra[45].

**5.2.3 Clasificación**. La clasificación es una tarea tradicional de minería de datos.

> Proponen un algoritmo de clasificación de vecinos más cercano, utilizando la escala apropiada derivada. Kadous y Sammut (2005) usan el enfoque de meta característica (es decir, la subestructura recurrente) como máximos locales en series de tiempo para generar clasificadores. Del mismo modo, Yang et al. (2005) se centran en la selección de subconjuntos de características (FSS) basados en componentes principales comunes, que se llama CleVer, para retener la información de correlación entre las características originales. La clasificación se emplea para evaluar la efectividad del subconjunto seleccionado de características.
>
> Se adoptan técnicas de agrupamiento para resumir y producir una descripción compacta de rasgos destacados y sus relaciones. Boyd (1998) desarrolló un sistema que integra el procesamiento de señales basado en el conocimiento y el procesamiento del lenguaje natural para generar automáticamente descripciones, y se prueba en los datos meteorológicos. Estas descripciones se basan en tendencias a corto y largo plazo, que se detectan utilizando la transformación wavelet. Guimaraes y Ultsch (1999) proponen una aproximación a patrones de tránsito en series temporales multivariadas a una descripción lingüística. Las reglas gramaticales temporales de diferentes niveles de abstracción se extraen de los resultados de redes neuronales y otros métodos exploratorios. Este enfoque se aplica a series temporales médicas, es decir, trastornos respiratorios relacionados con el sueño (Guimaraes et al., 2001) [46].

➢**Regresión.** La regresión crea modelos predictivos, la diferencia entre la regresión y clasificación es que la regresión tiene como objetivo atributos numéricos / continuos, mientras que la clasificación los atributos discretos / categoría. En otras palabras, si el atributo de destino contiene valores continuos (de coma flotante), se requiere una técnica de regresión. Si el atributo de destino contiene valores categóricos (cadena o un entero discreto), una técnica de clasificación se necesita.

La forma más común de la regresión es la regresión lineal, en el que se calcula una línea que mejor se ajusta a los datos, es decir, la línea que minimiza la distancia media de todos los puntos de la línea.

Esta línea se convierte en un modelo predictivo cuando no se conoce el valor de la variable dependiente; su valor está en relación con el punto de la línea que corresponde a los valores de las variables independientes para ese registro.

---

[45] FU, Op. cit., p. 164
[46] Ibíd., p.165

**5.2.4 Agrupación**. "Es la tarea de encontrar grupos de objetos similares o relacionados, de manera que dos objetos del mismo grupo son similares o relacionados entre sí y diferentes de los objetos de otros grupos"[47].

Agrupación de objetos y cosas en diferentes grupos es una forma común de describir el mundo, algunos dominios donde se utiliza ampliamente son:

- Biología: los biólogos siempre han intentado organizar todos los seres vivos en grupos basados en sus similitudes de estructura.

- Recuperación de información: la web contiene miles de millones de páginas web, de modo que una consulta en un motor de búsqueda puede devolver millones de resultados, estos resultados pueden agruparse en grupos, lo que facilita al usuario explorar los resultados de su consulta.

- Clima: encontrar patrones en la atmósfera y el océano ayudan a entender.

- Psicología y medicina: la agrupación puede usarse para determinar los diferentes tipos de enfermedad, y también puede usarse para encontrar patrones en su distribución temporal y espacial.

- Negocio: segmentar a los clientes en grupos para análisis adicional y marketing de destino.

Uno de los algoritmos de agrupamiento ampliamente utilizados es el algoritmo de K-means, que es un algoritmo simple y directo que realiza el clustering asignando vectores de un conjunto de datos dado a los clusters representados por un vector llamado centroide[48].

**5.2.5 Rapidminer**. RapidMiner es un entorno para el aprendizaje automático y para procesos de minería de datos, bajo el concepto de operador modular permite el diseño de cadenas de operadores complejos anidados para un gran número de problemas de aprendizaje. El manejo de los datos es transparente para los usuarios, ya no tienen que hacer frente con el formato de los datos reales o de los datos desde diferentes puntos de vista, el núcleo de RapidMiner se ocupa de las transformaciones necesarias.

Hoy en día, RapidMiner es el líder mundial en soluciones de Minería de Datos con código abierto y es "ampliamente utilizado por los investigadores y las empresas. RapidMiner introduce nuevos conceptos de manejo transparente de datos facilita el proceso de configuración para usuarios finales, además de que cuenta con

---

[47] EDDIB, A.J.A.; MOHAMMED, E.M. y CHAHHOU, M. Algorithms and systems for data mining: A survey. En: Colloquium in Information Science and Technology, CIST. Enero, 2015. no. 2, p. 107-114.
[48] Ibíd., p. 107.

interfaces claras y una especie de lenguaje de script basado en XML lo que la convierte en un entorno de desarrollo integrado para la Minería de Datos y el aprendizaje automático”[49].

[49] LAND, Sebastián y FISCHER, S. RapidMiner in academic use, V, 1-3 [en línea]. Berlín: Rapid-I GmbH [citado 20 agosto, 2017]. Disponible en Internet: <URL: http://docs.rapidminer.com/downloads/RapidMiner_RapidMinerInAcademicUse_en.pdf>

# 6. ENTENDIMIENTO DEL NEGOCIO

## 6.1 PROYECTO DE INVESTIGACIÓN

Se enmarca en un proyecto de investigación que tiene un tiempo estimado de dos años, donde participan los siguientes grupos de investigación:

- Grupo de Investigación en Derecho Público y TIC – Facultad de Derecho

- Grupo de Software Inteligente y convergencia Tecnológica GISIC – Facultad de Ingeniería

Las áreas transversales en las que se desarrolla el proyecto son:

- El área transversal de la persona y cultura que se enmarca en el tema Estado, Derecho y Sociedad.

- El área transversal de Gestión y Tecnología al servicio de la sociedad que se enmarcan dentro del tema Derecho público y TIC.

Se pretende una investigación a partir del método deductivo que permita estructurarlo por capítulos:

- Determinar problemas concretos desde una perspectiva ética y jurídica constitucional que genera el Big data.

- Concretar los problemas potenciales que se presienten entorno al uso del big data y su posible naturaleza jurídica.

- Implementar un prototipo de software que permita evaluar el grado de madurez de la implementación de la política de datos abiertos, sus principios y la calidad de los datos generados por las entidades públicas del gobierno generando un dato relativo por medio del re-procesamiento.

- Implementar un modelo estadístico que permita describir el comportamiento de los datos generados a partir de los contratos de las entidades públicas del gobierno, logrando la entrega de datos estadísticos descriptivos e inferenciales al usuario final.

Para el desarrollo de software se propone la construcción de una arquitectura web con el fin de garantizar que el usuario final pueda acceder desde cualquier lugar por medio de un navegador web. Algunos atributos de calidad que van a ser considerados durante el desarrollo del software son: escalabilidad, eficiencia y disponibilidad.

Se espera que la investigación tenga incidencia sobre sobre el desarrollo y

aplicación del Big data y su interpretación a medida que se aborde el tema en casos concretos. Se desea que el resultado sea tomado en cuenta por otros sistemas judiciales y en la adopción de políticas públicas debido a la colaboración internacional que recibirá el proyecto.

## 6.2 ENTENDIMIENTO DE LOS DATOS

El enfoque del proyecto está centrado en la plataforma SECOP (Sistema Electrónico de Contratación Pública) donde son registrados los procesos de compra pública por entidades Estatales, en esta se publican documentos del proceso desde la planeación del contrato hasta su liquidación. También permite a las Entidades Estatales y al sector privado tener una comunicación abierta y regulada sobre los Procesos de Contratación.

Existe una siguiente versión de SECOP (Sistema Electrónico de Contratación Pública) llamada SECOP II para pasar de la simple publicidad a una plataforma transaccional que permite a Compradores y Proveedores realizar el Proceso de Contratación en línea.

La actividad contractual de las entidades que ejecutan recursos públicos debe estar publicada en el SECOP de acuerdo con lo establecido en la Ley 1150 de 2007 y en el Decreto Ley 019 de 2012.

En los datos encontramos las siguientes Atributos:

- Área geográfica. El área geográfica hacer referencia a donde se ejecuta el contrato, se encuentra dividida en departamento y ciudad de Colombia.

- Costo: Los costos hacen relación a los recursos económicos necesarios para el proyecto.

- Tipo de Proceso. Los tipos de procesos hace referencia a como se realizó el contrato, por ejemplo: Contratación Mínima Cuantía, Contratación Directa (Ley 1150 de 2007) y Licitación Pública.

- Tipo de Contrato. Los tipos de contratos se refiere a la forma en que se va a desarrollar el contrato con determinados características y condiciones.
Encontramos: prestación de servicios, suministro, obra, compraventa, concesión, consultoría.

- Estado del contrato. El estado del contrato hace referencia a como se encuentra el contrato en un determinado momento. En este caso va a ser principalmente Celebrado.

- Fecha de liquidación. Hace referencia al momento en el que finalizo el contrato.

- Como datos Contratistas encontramos. Nombre de la entidad

- Como datos del Contratante encontramos. nombre de la entidad

- Detalle. Encontramos información adicional relacionada con el contrato.

- Fecha del contrato. Hace relación a la fecha cuando se firmó el contrato.

- Fecha de liquidación. Hace relación al momento que se da por finalizado el contrato.

- Fecha de terminación anticipada. Esta fecha la encontramos cuando el contrato no finalizo en la fecha establecida por algún motivo.

- Numero contrato. Identificador único del contrato.

- Numero de proceso. Identificador único de una determinada agrupación de contratos determinados características.

- Nivel de la entidad. Nivel de cobertura de la Entidad que hace el registro del proceso de compra pública.

# 7. PREPARACIÓN DE LOS DATOS

El proceso estándar de minería de datos involucra (1) entender el problema, (2) preparar las muestras de datos, (3) desarrollar el modelo, (4) aplicar el modelo en un conjunto de datos para ver cómo funciona el modelo en el mundo real y (5) despliegue a producción.

El proceso de extraer información de los datos es iterativo. Los pasos dentro del proceso de minería de datos no son lineales y tienen muchos bucles, saltar entre los pasos y a veces volver al primer paso para redefinir la instrucción del problema de minería de datos.

## 7.1 HERRAMIENTA DE EXTRACCIÓN

Dando continuidad a un proyecto de extracción de datos de la fuente datos.gov.co, se extraen los datos que van a ser utilizados en el proyecto. Los datos extraídos son guardados en una base de datos (NOSQL) mediante una interfaz gráfica. En www.datos.gov.co son identificados por un código único, el cual es necesario para utilizar la herramienta de extracción de datos (véase la Tabla 1).

**Cuadro 1. Conjuntos de Datos Utilizados**

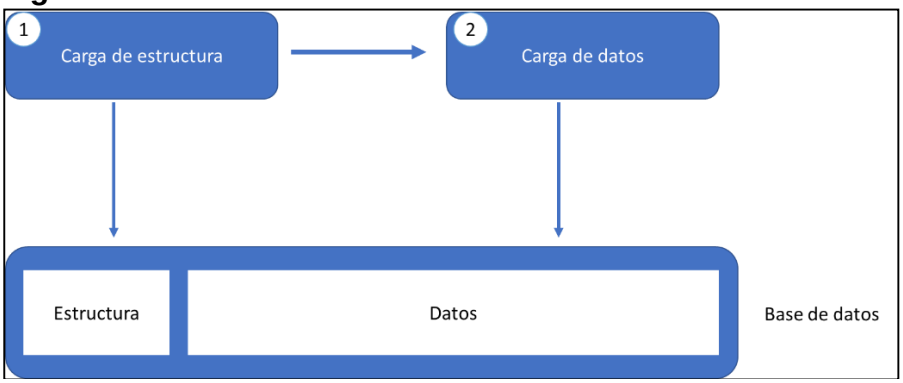
| Identificador | Nombre | Descripción |
|---|---|---|
| ewm2-yzgs | SECOP I - Consolidado | Procesos de compra pública registrados en la plataforma SECOP I |
| m58t-y685 | SECOP II - Consolidado | Procesos de compra pública registrados en la plataforma SECOP II |
| 4n4q-k399 | Multas-y-Sanciones-SECOP-I | Registro de las Multas y Sanciones generadas en la plataforma SECOP I |

Fuente. El Autor.

La estructura de la información en la base de datos guardada por la herramienta se divide en dos partes: metadatos y el conjunto de datos los cuales son guardados en formato JSON.

En el proceso de la herramienta de extracción se tienen dos pasos principales:

**Figura 4. Proceso Herramienta de Extracción**




Fuente. El Autor.

En la Figura 2 se observa que el proceso se compone de la carga de la estructura (1) del dataset para el cual se requiere el identificador único después de este paso se realiza el segundo paso de cargar los datos requiriendo como parámetro el identificador único. En la base de datos quedan guardadas las dos partes por separado para su manipulación.

## 7.2 LIMPIEZA DE LOS DATOS

Se analizaron los datos guardados en la base de datos(MongoDB), en los conjuntos de datos se encuentran valores inválidos que se deben eliminar o reemplazados por valores coherentes:

➢Conjunto de datos 1. Contratos públicos activos publicados en el secop:

✓0.3716% de inconsistencias de un total de 461268 registros. Campo 11
✓0.1147% de inconsistencias de un total de 461268 registros. Campo 13

➢De los 50000 registros extraídos 5371 no tienen el año de firma del contrato, inicio de ejecución y finalización del contrato.

➢Se encontró que 5468 filas tenían algún campo nulo. Después de eliminar las filas con campos nulos quedan 44532 filas.

➢Se eliminan los registros con valor del contrato en cero, aproximadamente un 3% del total de registros.

➢Los contratos en liquidación son aproximadamente un 24% del total de registros y en definición la liquidación se da cuando ambas partes se pronuncian sobre la ejecución de las prestaciones contractuales, como también respecto a los sucesos presentados durante su desarrollo.

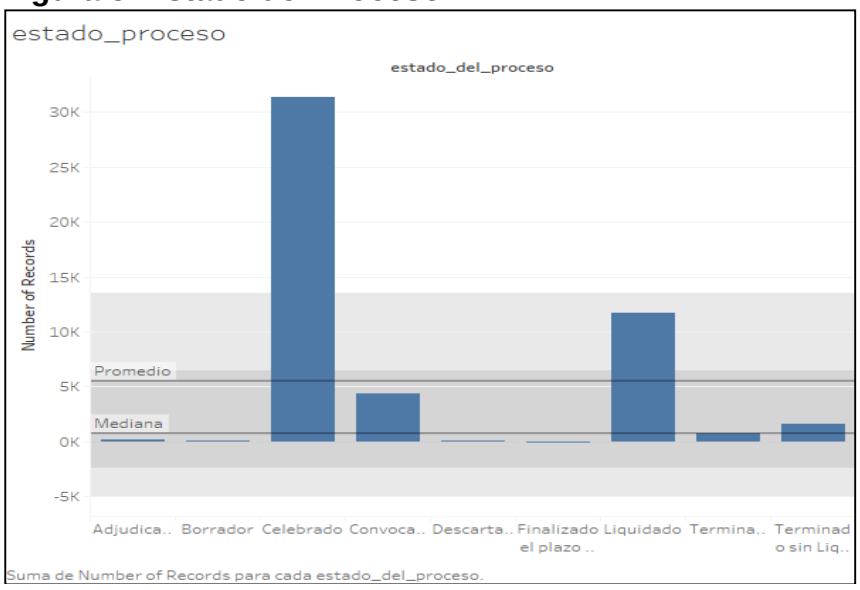
- Teniendo en cuenta lo anterior, se seleccionan del conjunto de datos de SECOP I los contratos firmados en el año 2016 y que son liquidados en ese mismo año. La calificación del contrato el 97% de 44532 registros no tienen definido una calificación, el restante que tiene datos no está estandarizado, se evidencian varias escalas.

- En el conjunto de datos Multas-y-Sanciones-SECOP-I no tiene registro, se esperaba realizar un cruce con el conjunto de datos de SECOP I.

## 8. EXPLORACIÓN DE LOS DATOS

En la exploración de los datos se realiza un análisis descriptivo con el objetivo de escoger las variables a utilizar en el proceso de minería de datos (véase las Figuras 5, 6, 7, 8, 9 y 10).

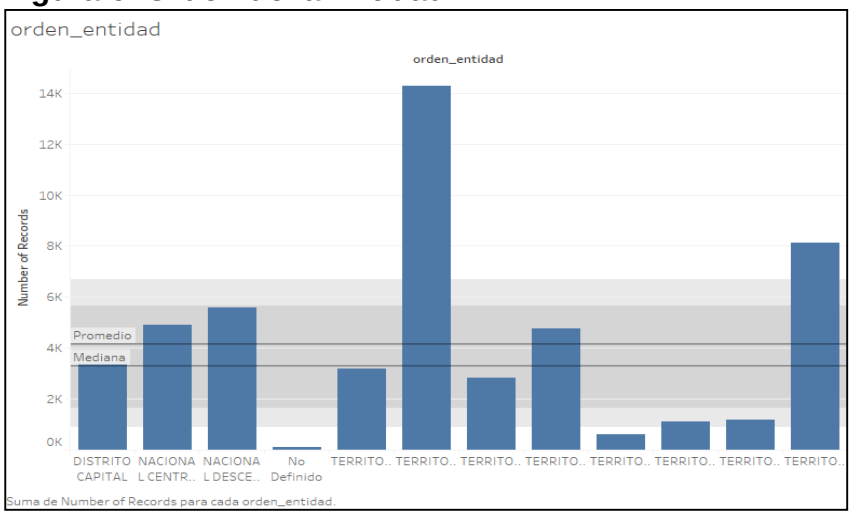
**Figura 5. Estado del Proceso**




Fuente. El Autor

En la Figura 5 aproximadamente el 60% de los contratos están celebrado y un 12% Liquidados, los contratos que hayan terminado o liquidado no tendrán valores para los días de ejecución.

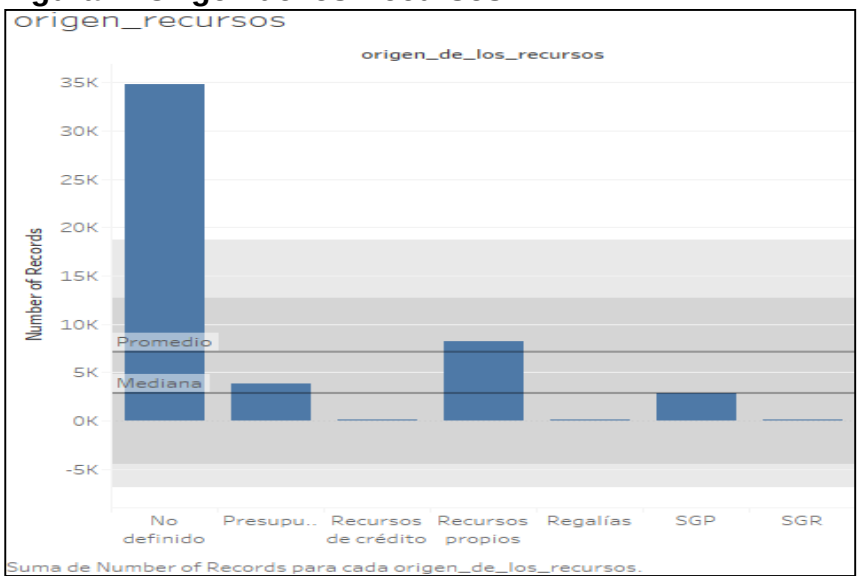
**Figura 6. Orden de la Entidad**




Fuente. El Autor

En la Figura 6 se observa que el orden de las entidades se encuentra distribuido entre los distintos valores, siendo una posible variable a tener en cuenta.

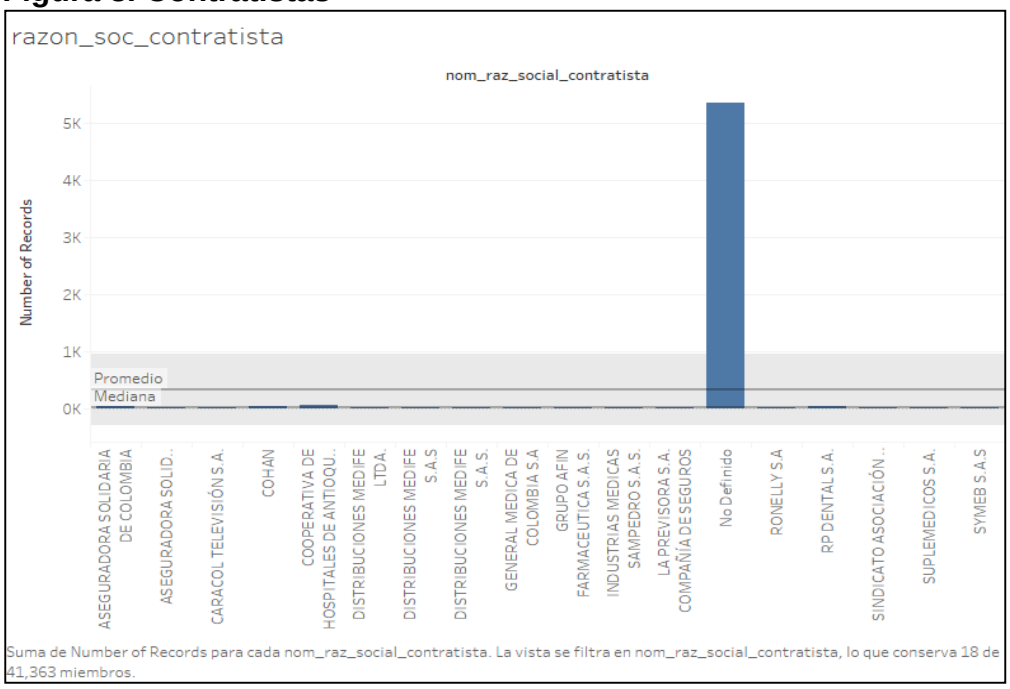
**Figura 7. Origen de los Recursos**




Fuente. El Autor

En la Figura 7 aproximadamente el 70% no tiene definido el origen de los recursos, por lo tanto, esta variable no se tendrá en cuenta.

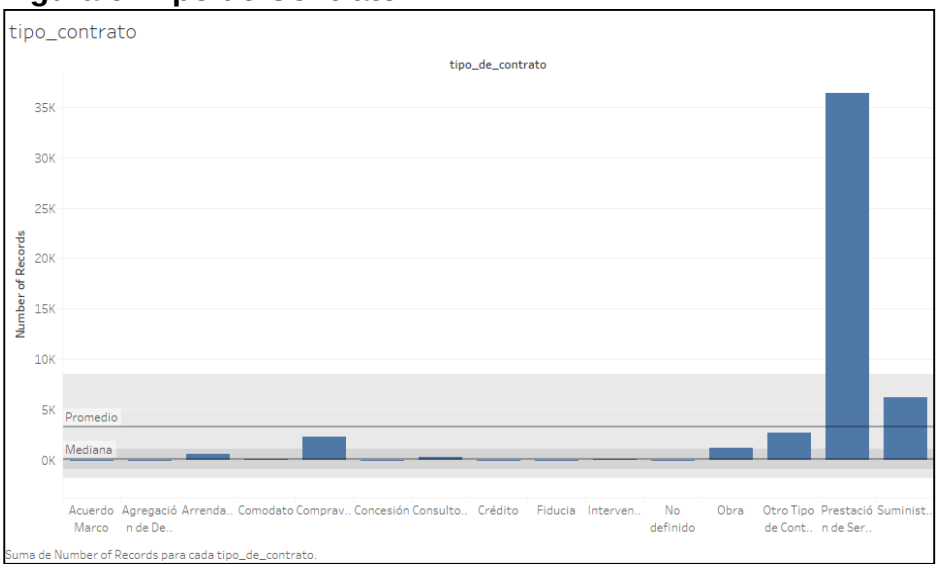
**Figura 8. Contratistas**




Fuente. El Autor

Más del 80% de los contratos no tienen definido el nombre o razón social del contratista, se debe filtran los contratos que no tienen contratista.

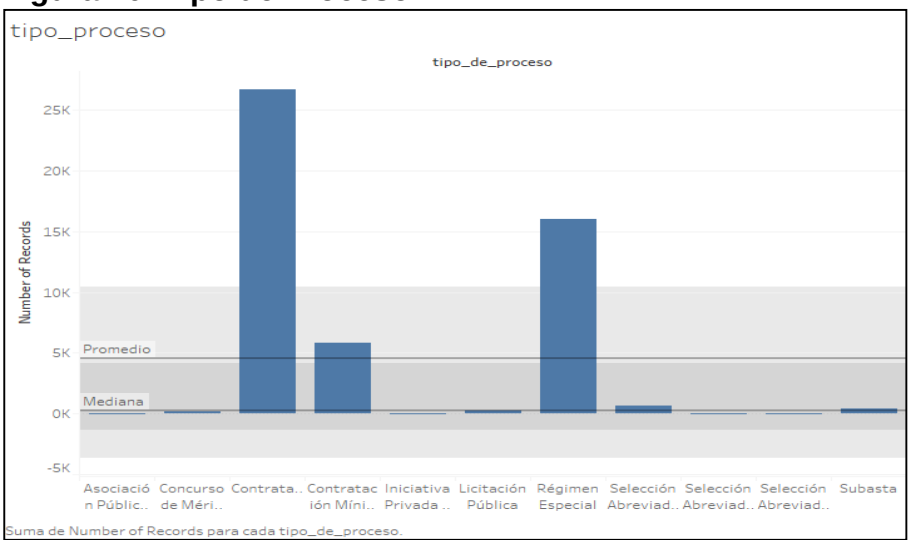
**Figura 9. Tipo de Contrato**




Fuente. El Autor

En la Figura 9 más del 50% de los contratos son por prestación de servicios, por lo tanto, es una modalidad que prima en el sector público.

**Figura 10. Tipo de Proceso**




Fuente. El Autor

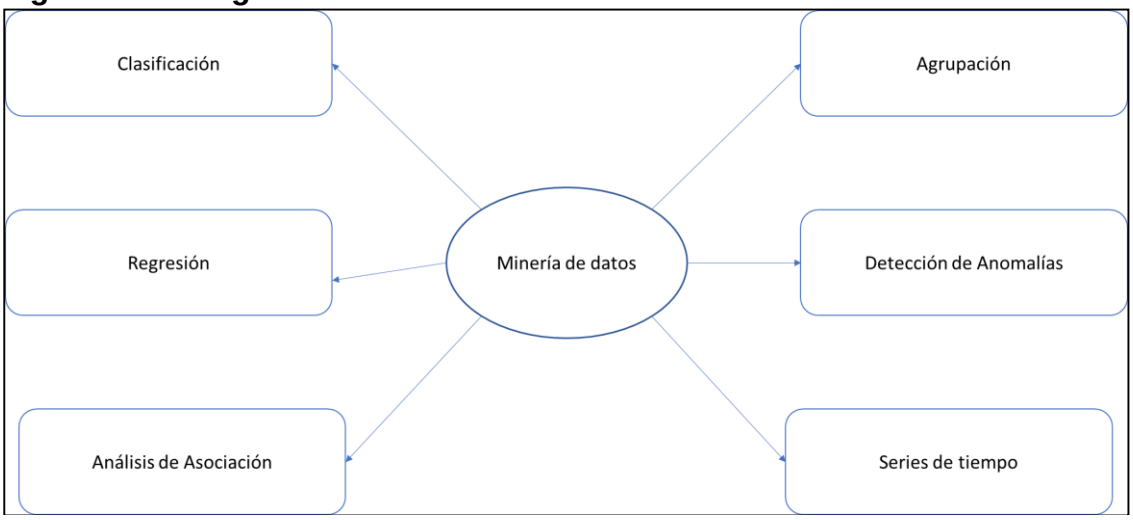
En la Figura 10 aproximadamente el 50% de los contratos tienen el tipo de proceso de contratación Directa, el otro 50% se divide principalmente en Contratación de Mínima Cuantía y Régimen Especial.

## 9. ANÁLISIS DE MODELOS

En minería de datos se encuentran tareas que hacen uso de una variedad de algoritmos. Basándose en el problema de los datos, la minería de datos se clasifica en tareas tales como clasificación, análisis de asociación, agrupación y regresión.

Cada tarea de minería de datos utiliza algoritmos específicos como por ejemplo árboles de decisión, redes neuronales, k-vecinos más cercanos, k-medios de agrupación (véase la Figura 11).

**Figura 11. Categorías Minería de Datos**




Fuente. El Autor

En la Figura 11 se observa como los problemas de minería de datos también pueden agruparse en clasificación, regresión, análisis de asociación, detección de anomalías, series temporales y tareas de minería de texto.

A continuación, se listan las categorías de los algoritmos que son utilizados en minería de datos:

**Cuadro 2. Descripción Categorías Minería de Dato**

| Categoría | Descripción | Algoritmos | Ejemplos |
|---|---|---|---|
| Clasificación | Predice si un dato (data point) pertenece a una de las clases predefinidas. La predicción se basará en el aprendizaje de un conjunto de datos conocidos. | Árboles de decisión, Redes neuronales, Modelos bayesianos, Reglas de inducción, K vecinos más cercanos | Asignación de los votantes en los cubos conocidos por los partidos políticos, por ejemplo: las mamás de fútbol. Recoger nuevos clientes en uno de los grupos de clientes conocidos. |
| Regresión | Predecir la etiqueta numérica de destino de un punto de datos. La predicción se basará en el aprendizaje de conjuntos de datos conocidos. | Regresión lineal, Regresión logística | Predicción de la tasa de desempleo para el próximo año. Estimación de la prima del seguro. |
| Detección de anomalías | Predecir si un punto de datos es un valor atípico en comparación con otros puntos de datos del conjunto de datos. | Basado en la distancia, basado en la densidad, LOF | Detección de transacciones fraudulentas en tarjetas de crédito. Detección de intrusiones de red. |
| Series de tiempo | Predecir si el valor de la variable de destino para el marco de tiempo futuro basado en los valores del históricos. | Suavizado exponencial, ARIMA, regresión | Pronóstico de ventas, pronóstico de producción, prácticamente cualquier fenómeno de crecimiento que deba extrapolarse. |
| Agrupación | Identificar agrupaciones naturales dentro del conjunto de datos basado en propiedades heredadas dentro del conjunto de datos. | K means, Agrupación basado en densidad - DBSCAN | Encontrar segmentos de clientes en una empresa basados en datos de llamadas de transacciones, web y clientes. |
| Análisis de la asociación | Identificar las relaciones dentro de un conjunto de artículos basado en datos de transacciones. | FP Growth, Apriori | Encontrar oportunidades de venta cruzada para un minorista basado en el historial de compras de transacciones. |

Fuente. El Autor

A continuación, se describe cada uno de los algoritmos relevantes que se encuentran cada una de las tareas descritas anteriormente:

## 9.1 CLASIFICACIÓN

➢Decision Trees (Ver sección 0). Particiona los datos en subconjuntos más pequeños donde cada subconjunto contiene (en su mayoría) respuestas de una clase ('si' o 'no') con la posibilidad de expresarlo como conjuntos de reglas. Son bastante eficientes y obtienen resultados para clasificación.

➢Rule Induction (Ver sección 0). Modelos de la relación entre la entrada y la salida mediante la deducción simple IF / THEN reglas de un conjunto de datos.

➢k-Nearest Neighbors (Ver sección 0). Un aprendiz perezoso donde ningún modelo es generalizado. Cualquier punto nuevo de datos desconocidos se compara con puntos de datos conocidos similares en el conjunto de entrenamiento.

➢Naïve Bayesian (Ver sección 0). Predice la clase de salida basada en el teorema de Bayes calculando la probabilidad condicional de la clase y la probabilidad previa.

➢Ensemble Learners (Ver sección 0). Aprovecha la sabiduría de la multitud. Emplea una serie de modelos independientes para hacer una predicción y agrega la predicción final.

## 9.2 REGRESIÓN

➢Linear Regression (Ver sección 0). El modelo predictivo clásico que expresa la relación entre las entradas y un parámetro de salida en forma de una ecuación.

➢Logistic Regression (Ver sección 0). Técnicamente, este es un método de clasificación. Pero estructuralmente es similar a la regresión lineal.

## 9.3 ANÁLISIS DE ASOCIACIÓN

➢Apriori (Ver sección 0) y FP-Growth (Ver sección 0). Mide la fuerza de la co-ocurrencia entre un artículo con otro.

## 9.4 AGRUPACIÓN

➢k-Means (Ver sección 0). Conjunto de datos se divide en k clusters por encontrar k centroides.

➢DBSCAN (Ver seccion 0). Identifica las agrupaciones como un área de alta densidad rodeada por áreas de baja densidad.

- Self-Organizing Maps (Ver sección 0). Una técnica de agrupamiento visual con raíces de redes neuronales y clustering de prototipos.

## 10. EVALUACIÓN DE LOS MODELOS

La elección del algoritmo a utilizar depende principalmente del tipo de conjunto de datos, el objetivo de la minería de datos, la estructura de los datos, la presencia de valores atípicos, la potencia computacional disponible, el número de registros, el número de atributo. Corresponde al profesional de la minería de datos tomar una decisión sobre qué algoritmo utilizar evaluando el rendimiento de varios algoritmos.

En la actualidad se encuentran cientos de algoritmos desarrollados en las últimas décadas para resolver problemas de minería de datos, los cuales se evaluarán con el objetivo de escoger el que mejor se adapte a este caso.

A continuación, se describen los criterios de selección encontrados en el Cuadro 3:

- Conjunto de datos. El conjunto de datos hace referencia a los tipos de datos que se pueden encontrar, por ejemplo, datos cuantitativos que son valores numéricos o cualitativos que son valores no numéricos.

- Estructura. Hace referencia a como están organizados los datos. por ejemplo, si la estructura es jerárquica (json) o una matriz de datos (tabla o dataset).

- No requiere conjunto de datos de entrenamiento. Hace alusión a si el algoritmo es supervisado o no supervisado, es decir, si se requiere datos previos para realizar el análisis como por ejemplo para realizar predicción es necesario datos previos.

- Número de registros. Hace alusión al comportamiento que tiene cada algoritmo con un volumen alto de datos, se relaciona con el tiempo que le toma realizar el proceso al algoritmo.

- Objetivo. El objetivo que se tiene es encontrar datos desconocidos, aunque encontramos otros enfoques como lo son la descripción o predicción de datos.

- Anomalía en los datos. Hace referencia a cuando se tiene ruido en el conjunto de datos, al comportamiento que tiene el algoritmo en estos contextos. Algoritmos como DBSCAN maneja de forma aceptable las inconsistencias en los datos.

**Cuadro 3. Evaluación de Algoritmos**

| Categoría | Algoritmos | Conjunto de datos | Estructura | No requiere Conjuntos de datos de entrenamiento | Numero de registros | Numero de atributos | Objetivo | Anomalias en los datos | Total |
|---|---|---|---|---|---|---|---|---|---|
| Clasificación | Decision Trees | | ✓ | | ✓ | ✓ | | | 3 |
| | Neural networks | ✓ | ✓ | | ✓ | ✓ | | | 4 |
| | Bayesian models | ✓ | ✓ | | ✓ | ✓ | | | 4 |
| | Induction rules | ✓ | | | ✓ | ✓ | | | 3 |
| | K nearest neighbors | ✓ | | | ✓ | ✓ | | | 3 |
| Regresión | Linear regression | | ✓ | | ✓ | ✓ | | | 3 |
| | Logistic regression | ✓ | ✓ | | | ✓ | | | 3 |
| Agrupación | K means | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | 6 |
| | DBSCAN | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 7 |
| Analisis de Asociación | FP Growth | ✓ | | ✓ | ✓ | ✓ | | | 4 |
| | Apriori | ✓ | | ✓ | ✓ | ✓ | | | 4 |

Fuente. El Autor

## 10.1 RESULTADO

Los algoritmos clasificados en Agrupación se acomodan al proyecto debido a que su objetivo es aglomerar un conjunto de objetos para encontrar si existe alguna relación entre ellos. Este tipo de algoritmos son no supervisados, es decir, no requieren un conjunto de datos de entrenamiento.

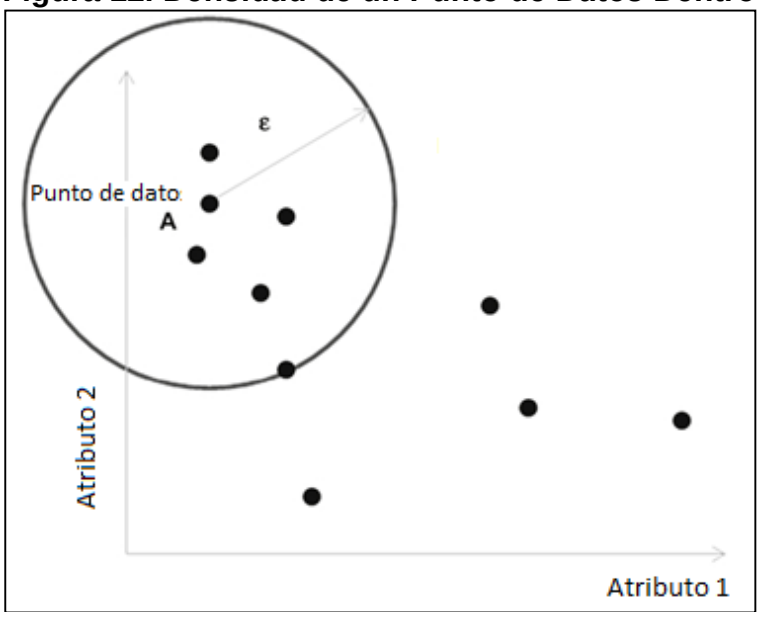
El algoritmo seleccionado: DBSCAN.
Funcionamiento del algoritmo de agrupación DBSCAN:

La densidad se puede definir como el número de puntos de datos en un espacio unitario n-dimensional. El número de dimensiones n es el número de atributos en un conjunto de datos. Técnicamente, la densidad se refiere al número de puntos en el espacio unitario, en este caso un cuadrante. Donde quiera que haya espacio de alta densidad entre los espacios de relativamente baja densidad, hay un grupo.

El algoritmo DBSCAN crea agrupaciones identificando el espacio de alta densidad y baja densidad dentro del conjunto de datos. De forma similar a la agrupación de k-means, se prefiere que los atributos sean numéricos porque el cálculo de distancia se sigue utilizando. Podemos reducir el algoritmo a tres pasos: definición de densidad umbral, clasificación de puntos de datos y agrupación.

**10.1.1 Definición de densidad umbral**. A continuación, se puede observar Densidad de un punto de datos dentro del radio ε. (véase la Figura 12).

**Figura 12. Densidad de un Punto de Datos Dentro del Radio ε.**




Fuente. El Autor

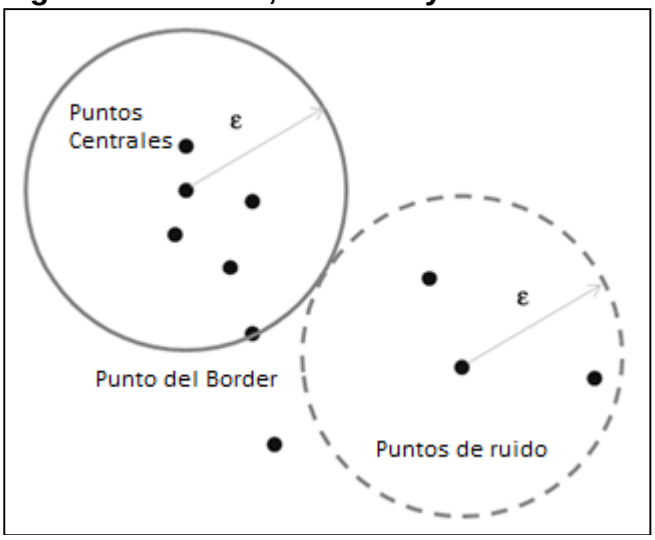
**10.1.2 Clasificación de puntos de datos**. En un conjunto de datos, con un ε dado y MinPoints, podemos clasificar todos los puntos de datos en tres cubos:

➢Puntos centrales. Todos los puntos de datos dentro de la región de alta densidad de al menos un punto de datos se consideran un punto central. Una región de alta densidad es un espacio donde hay al menos puntos de datos MinPoints dentro de un radio de ε para cualquier punto de datos.

➢Punto de borde. Los puntos de borde se sitúan en la circunferencia del radio ε desde un punto de datos. Un punto fronterizo es el límite entre el espacio de alta densidad y el de baja densidad. Los puntos fronterizos se cuentan dentro del cálculo de espacio de alta densidad.

➢Punto de ruido. Cualquier punto que no sea ni un punto central ni un punto de borde se llama punto de ruido. Forman una región de baja densidad alrededor de la región de alta densidad (véase la Figura 13).

**Figura 13. Núcleo, Frontera y Puntos de Densidad**




Fuente. El Autor

**10.1.3 Agrupación**. Una vez que todos los puntos de datos del conjunto de datos se clasifican en puntos de densidad, el agrupamiento es una tarea sencilla. Los grupos de puntos básicos forman grupos distintos. Si dos puntos centrales están dentro de ε y cada uno está dentro del otro, entonces ambos puntos centrales están dentro del mismo grupo.

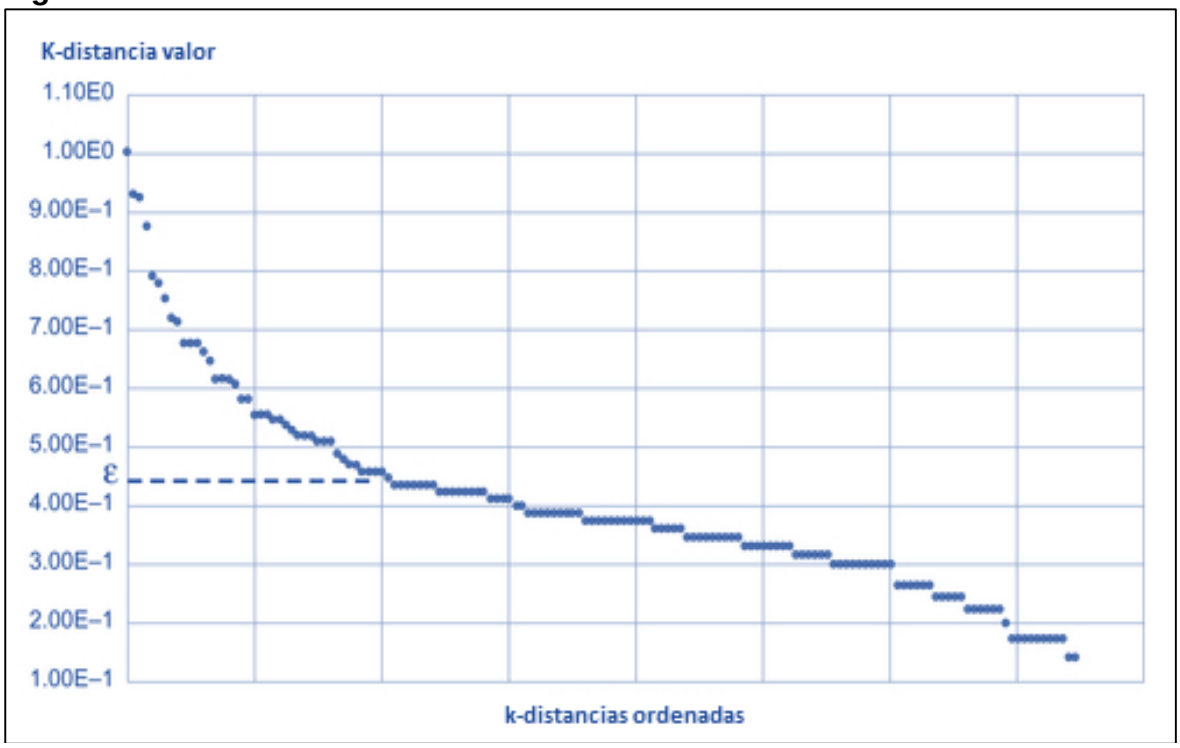
Todos estos núcleos agrupados forman un grupo, rodeado de puntos de ruido de baja densidad. Todos los puntos de ruido forman regiones de baja densidad alrededor del grupo de alta densidad, y los puntos de ruido no se clasifican en ningún grupo. Como DBSCAN es un algoritmo de agrupación parcial, algunos puntos de

datos se dejan sin etiquetar o asociados a un grupo de ruidos predeterminado.

**10.1.4 Optimización de Parámetros**. Una de las ventajas clave en el uso de un algoritmo de densidad es que no hay necesidad de especificar el número de clusters (k). Los clústeres se encuentran automáticamente en el conjunto de datos. Sin embargo, hay un problema de seleccionar el parámetro de distancia ε y un umbral mínimo (MinPoints) para identificar la región densa. Una de las técnicas utilizadas para estimar parámetros óptimos para el algoritmo de agrupación DBSCAN se refiere al algoritmo k-vecinos más cercano.

Podemos estimar los valores iniciales del parámetro construyendo un gráfico de distribución k. Para un valor especificado por el usuario de k (por ejemplo, cuatro puntos de datos), podemos calcular la distancia del k-ésimo vecino más cercano para un punto de datos. Si el punto de datos es un punto central en una región de alta densidad, entonces la distancia del k-ésimo vecino más cercano será menor.
Para un punto de ruido, la distancia será mayor. Del mismo modo, podemos calcular la distancia k para todos los puntos de datos en un conjunto de datos. Se puede construir un gráfico de distribución de k-distancias ordenando todos los valores de k-distancia de puntos de datos individuales en orden descendente como se observa en la Figura 14.

**Figura 14. Cálculo de Parámetro e**




Fuente. El Autor

El enfoque básico es observar el comportamiento de la distancia desde un punto hasta su vecino k más cercano, que vamos a nombrar k-distancia. Para los puntos

que pertenecen a algún grupo, el valor de k-distancia será pequeño si k no es mayor que el tamaño del clúster. Tenga en cuenta que habrá alguna variación, dependiendo de la densidad del grupo y la distribución aleatoria de los puntos, pero en promedio, el rango de variación no será enorme si las densidades de racimo no son radicalmente diferentes. Por otro lado, para los puntos que no están en un grupo, como los puntos de ruido, el k-distancia será relativamente grande. Por lo tanto, si calculamos la k-distancia para todos los puntos de datos para unos k, los clasificamos en orden creciente y luego trazamos los valores ordenados, esperamos ver un cambio brusco en el valor de k-distancia que corresponde a un valor adecuado valor de ε.

Si seleccionamos esta distancia como el parámetro ε y tomamos el valor de k como el parámetro MinPts, entonces los puntos para los cuales k-distacia es menor que ε serán etiquetados como puntos centrales, mientras que otros puntos serán etiquetados como ruido o puntos fronterizos.

**10.1.5 Algoritmo k-vecinos más cercano**. El algoritmo k-vecinos más cercano se basa en el aprendizaje por analogía, es decir, comparando una tupla de prueba dada con tuplas de entrenamiento que son similares a ella. Las tuplas de entrenamiento se describen mediante n atributos. Cada tupla representa un punto en un espacio n-dimensional. De esta manera, todas las tuplas de entrenamiento se almacenan en un espacio de patrones n-dimensional. Cuando se le da una tupla desconocida, un clasificador k-más cercano-vecino busca el espacio de patrón para las k tuplas de entrenamiento que están más cercanas a la tupla desconocida. Estas k tuplas de entrenamiento son los k "vecinos más cercanos" de la tupla desconocida.

La "Cercanía" se define en términos de una métrica de distancia, como la distancia euclidiana. La distancia euclidiana entre dos puntos o tuplas, siendo, X1 = (x11, x12 … x1n) y X2 = (x21, x22, …, x2n), es

Fórmula 1. Cálculo de distancias

$$\text{distacia}(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (X_{1i} - X_{2i})^2}$$

Fuente. El Autor

Para cada atributo numérico, tomamos la diferencia entre los valores correspondientes de ese atributo en la tupla X1 y en la tupla X2, cuadramos esta diferencia y la acumulamos. Se toma la raíz cuadrada del recuento de distancia acumulada total. Normalmente, normalizamos los valores de cada atributo antes de usar la fórmula 2. Esto ayuda a evitar que los atributos con rangos inicialmente grandes (como los ingresos) compensen los atributos con rangos inicialmente más pequeños (como los atributos binarios). Por ejemplo, la normalización mín-máx

puede utilizarse para transformar un valor v de un atributo numérico A en $v'$ en el rango [0, 1] calculando:

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

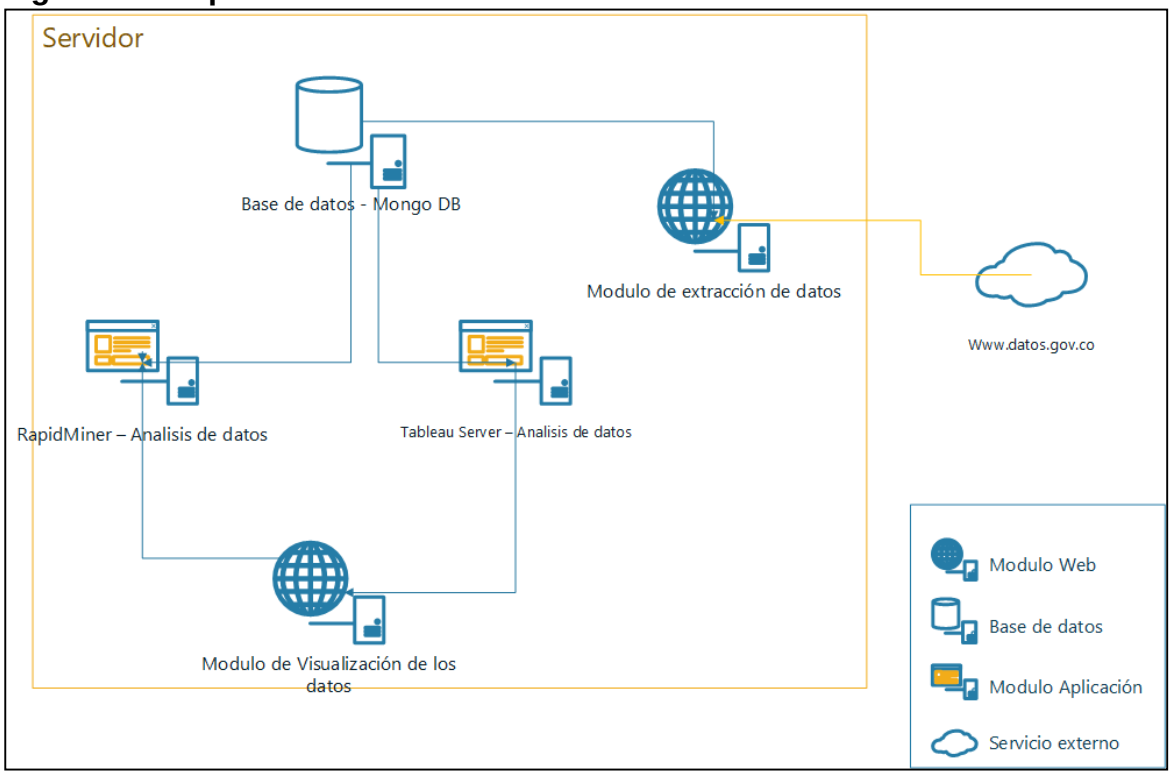
donde $\min_A$ y $\max_A$son los valores mínimo y máximo del atributo A.

Para k-más cercano-vecino, la tupla desconocida se asigna a la clase más común entre sus k vecinos más cercanos. Cuando k = 1, a la tupla desconocida se le asigna la clase de la tupla de entrenamiento que está más cerca de ella en el espacio del patrón. Los clasificadores vecinos más cercanos también se pueden usar para la predicción, es decir, para devolver una predicción de valor real para una tupla desconocida dada. En este caso, el clasificador devuelve el valor promedio de las etiquetas de valor real asociadas con los k vecinos más próximos de la tupla desconocida.

## 11. ARQUITECTURA DE SOFTWARE

A continuación, se realiza una descripción general de la arquitectura de la aplicación (véase la Figura 15).

**Figura 15. Arquitectura Global**




Fuente. El Autor

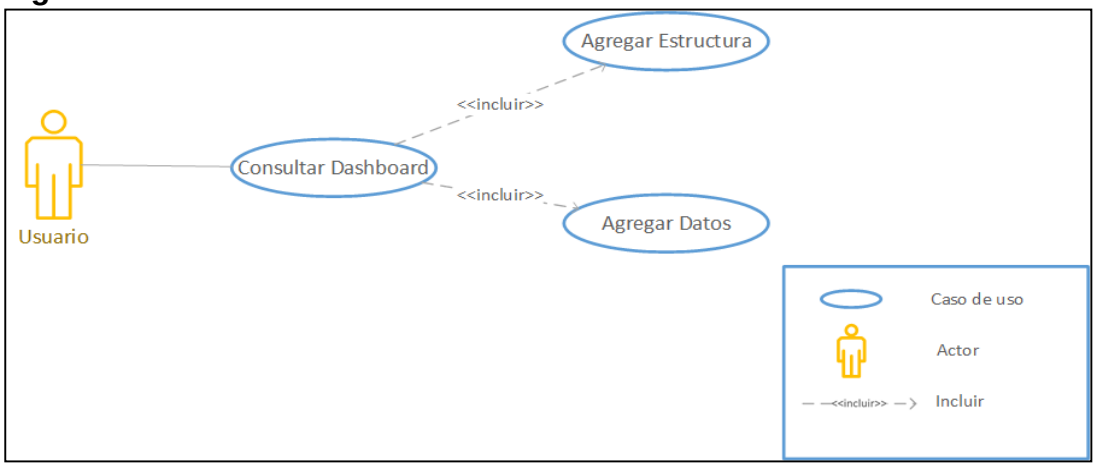
Los datos se extraen del servicio de datos abiertos (www.datos.gov.co) por medio del módulo de extracción, estos serán guardados en la base de datos mongoDB. La información guardada en la base de datos será consumida por el módulo de RapidMiner y Tableau para generar el análisis y posteriormente publicarlos en el dashboard.

A continuación, se describen las vistas de arquitectura de software desde el enfoque del autor Rozanski.

### 11.1 DESARROLLO

El usuario consultará los datos publicados en el dashboard por medio de un navegador web pero antes de realizar la consulta debe agregar la estructura de los datos y agregar datos, estas son funciones del módulo de extracción de datos (véase la Figura 16).

**Figura 16. Caso de Uso**

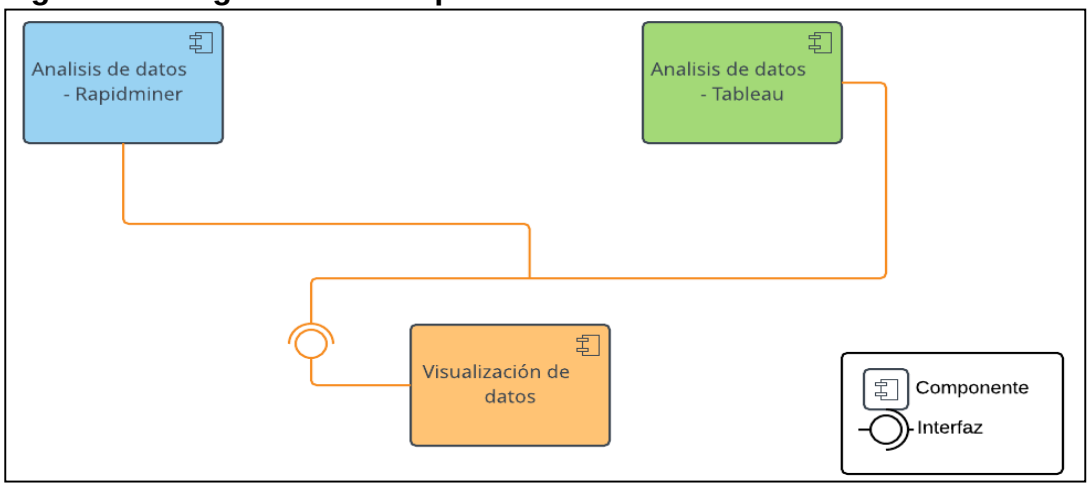



Fuente. El Autor

## 11.2 FUNCIONAL

Los componentes de Análisis de los datos – RapidMiner y Análisis de los datos – Tableau son los encargados de procesar los datos y el componente de Visualización de datos muestra el resultado de los otros dos componentes nombrados (véase la Figura 17).

**Figura 17. Diagrama de Componentes**

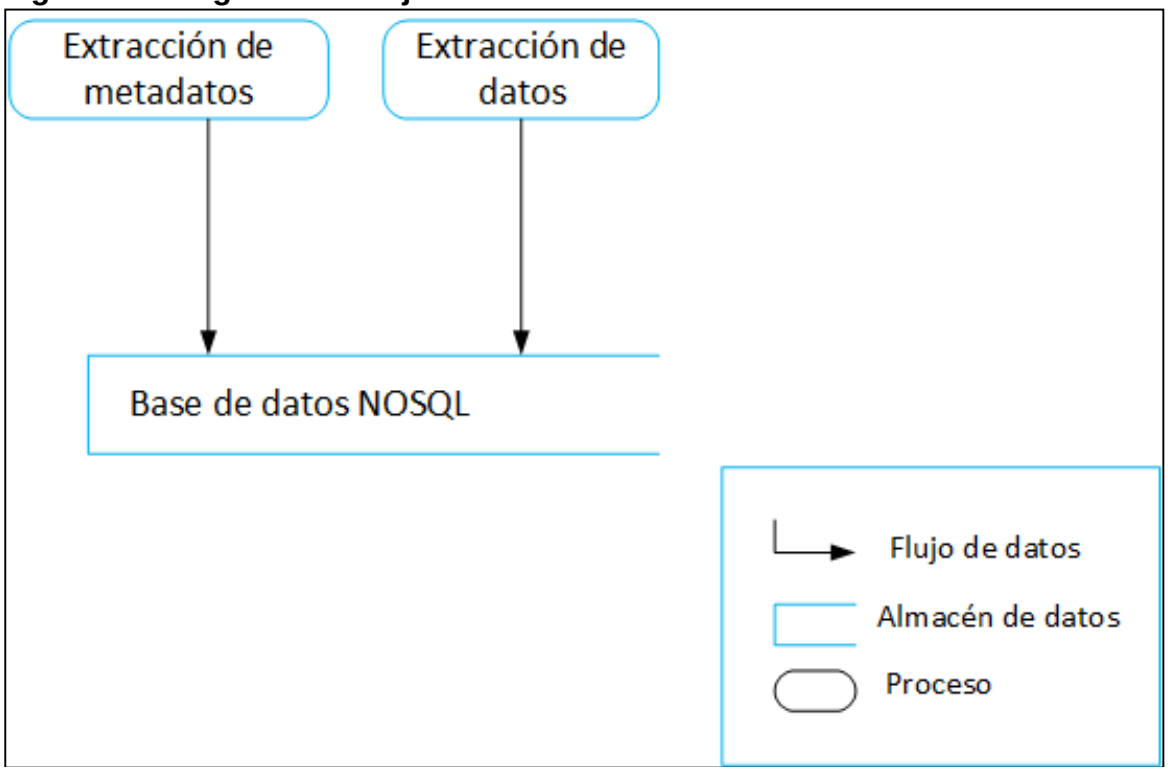



Fuente. El Autor

## 11.3 INFORMACIÓN Y DATOS

Al ejecutarse el proceso de extracción de datos se guardarán en la base de datos MongoDB, este proceso se encuentra dividido en dos partes: extracción de metadatos y extracción de los datos. Después de tener almacenado la información se realizará los procesos de análisis (véase la Figura 18).

**Figura 18. Diagrama de flujo de datos**

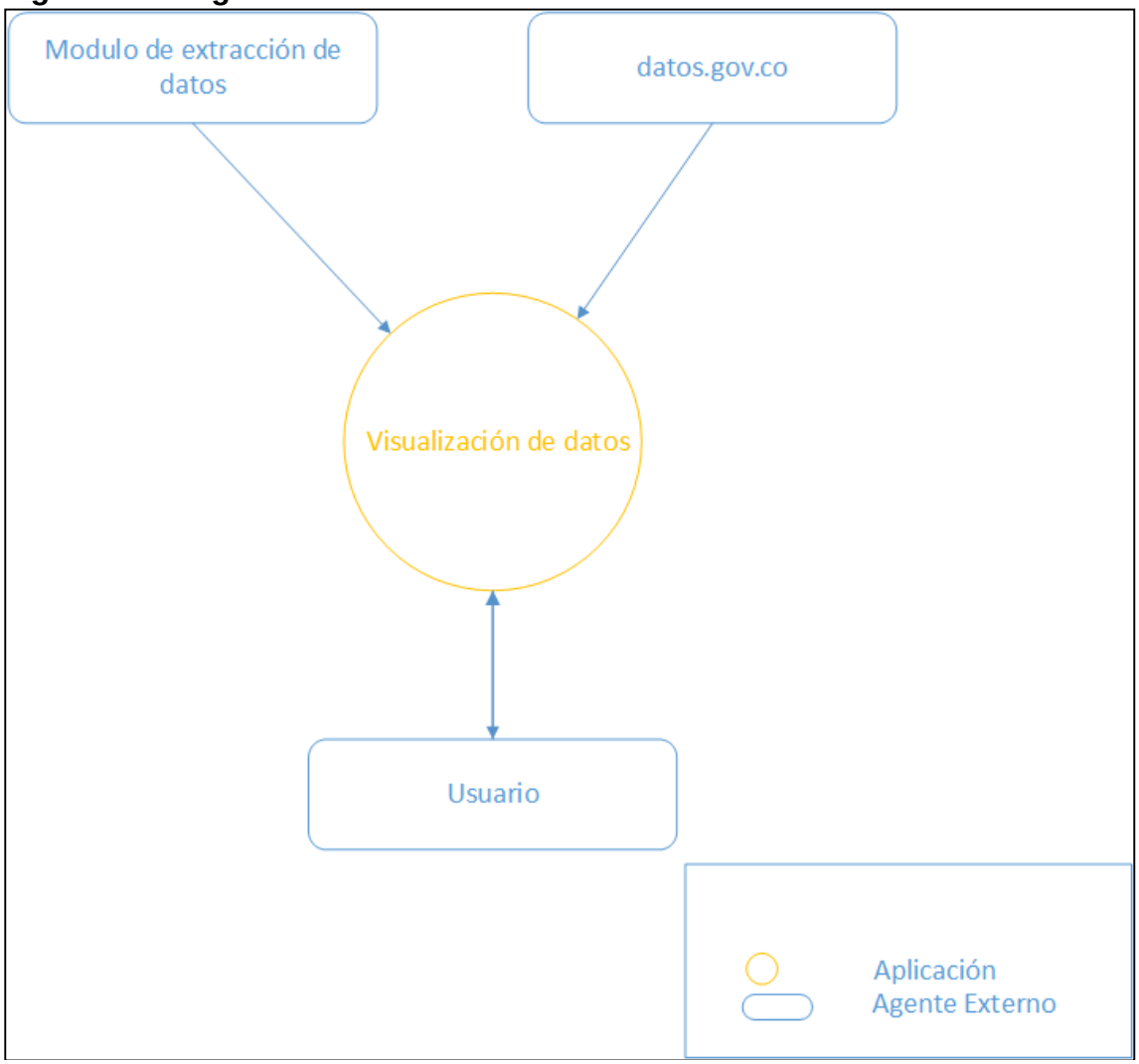



Fuente. El Autor

### 11.4 CONTEXTO

En la Figura 19 se observa una dependencia del componente de extracción de datos, quien es el encargado de guardar la información de la fuente www.datos.gov.co en la base datos con el propósito de realizar los procesos de análisis de datos y su posterior visualización (véase la Figura 19).

**Figura 19. Diagrama de Contexto**




Fuente. El Autor

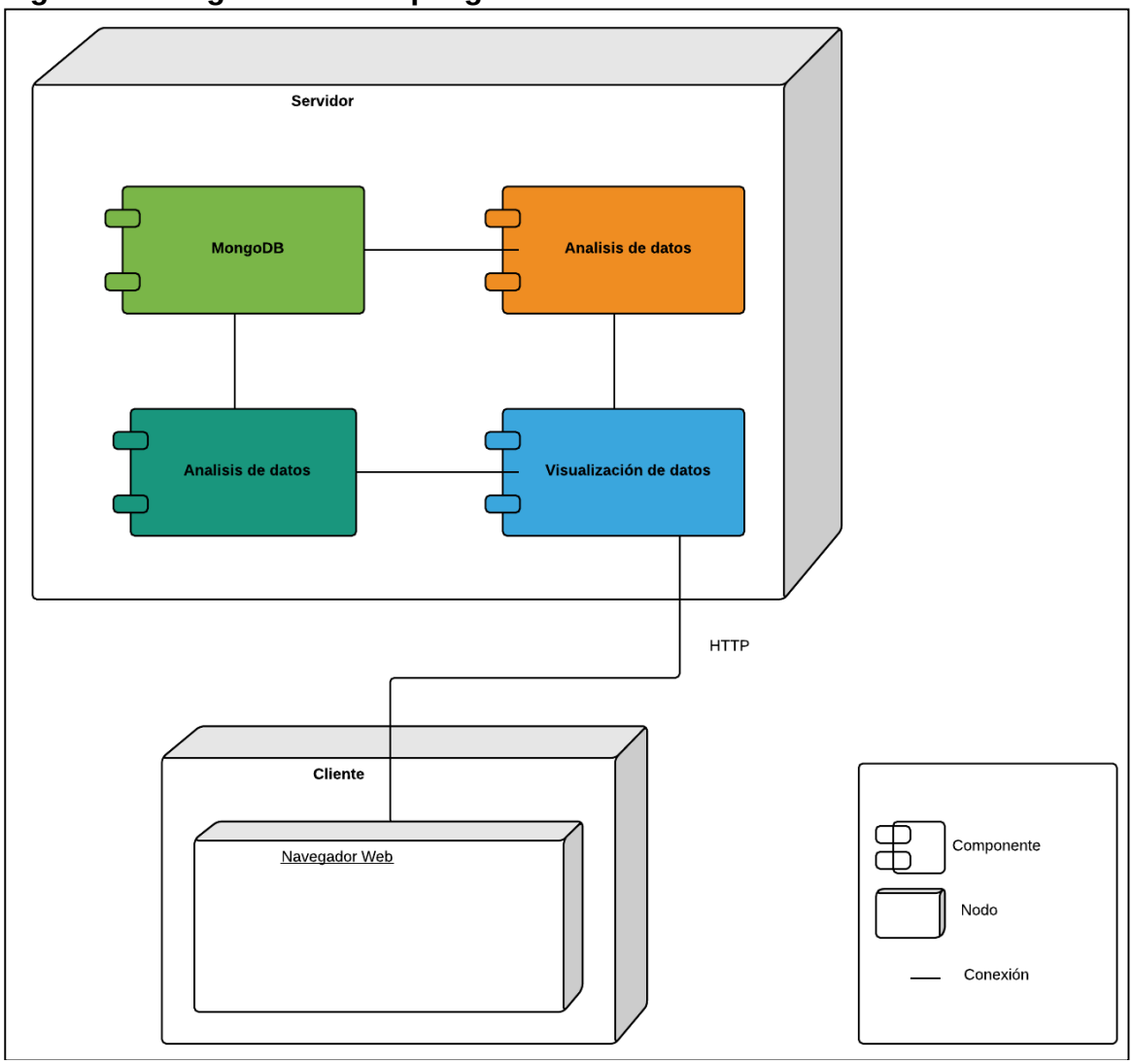
### 11.5 DESPLIEGUE

El servido se encuentra en las instalaciones de la universidad – Sede claustro. El sistema operativo es Windows 7, se encuentra virtualizado con las siguientes características:

- 6GB de memoria RAM
- 50GB de espacio de almacenamiento
- Procesador Xeon

Todos los componentes se ejecutarán desde un solo servidor, el componente de Visualización interactúa con los componentes de análisis de los datos por medio de peticiones HTTP (véase la Figura 20).

**Figura 20. Diagrama de Despliegue**




Fuente. El Autor

## 11.6 REQUERIMIENTOS NO FUNCIONALES

➢Dando continuidad al proyecto, encontramos que lo construido se encuentra implementado en Windows por lo tanto debe ser compatible con él. La aplicación se debe conectar a la base de datos mongodb y el desarrollo necesario deberá ser construido en Java EE.

➢El sistema debe tener una disponibilidad del 99,99% de las veces en que un usuario intente accederlo. (Disponibilidad)

➢La tasa de tiempos de falla del sistema no podrá ser mayor al 0,5% del tiempo de operación total. (Fiabilidad)

➢La interfaz de usuario será implementada para navegadores web únicamente con HTML5 y JavaScript. (Restricción)

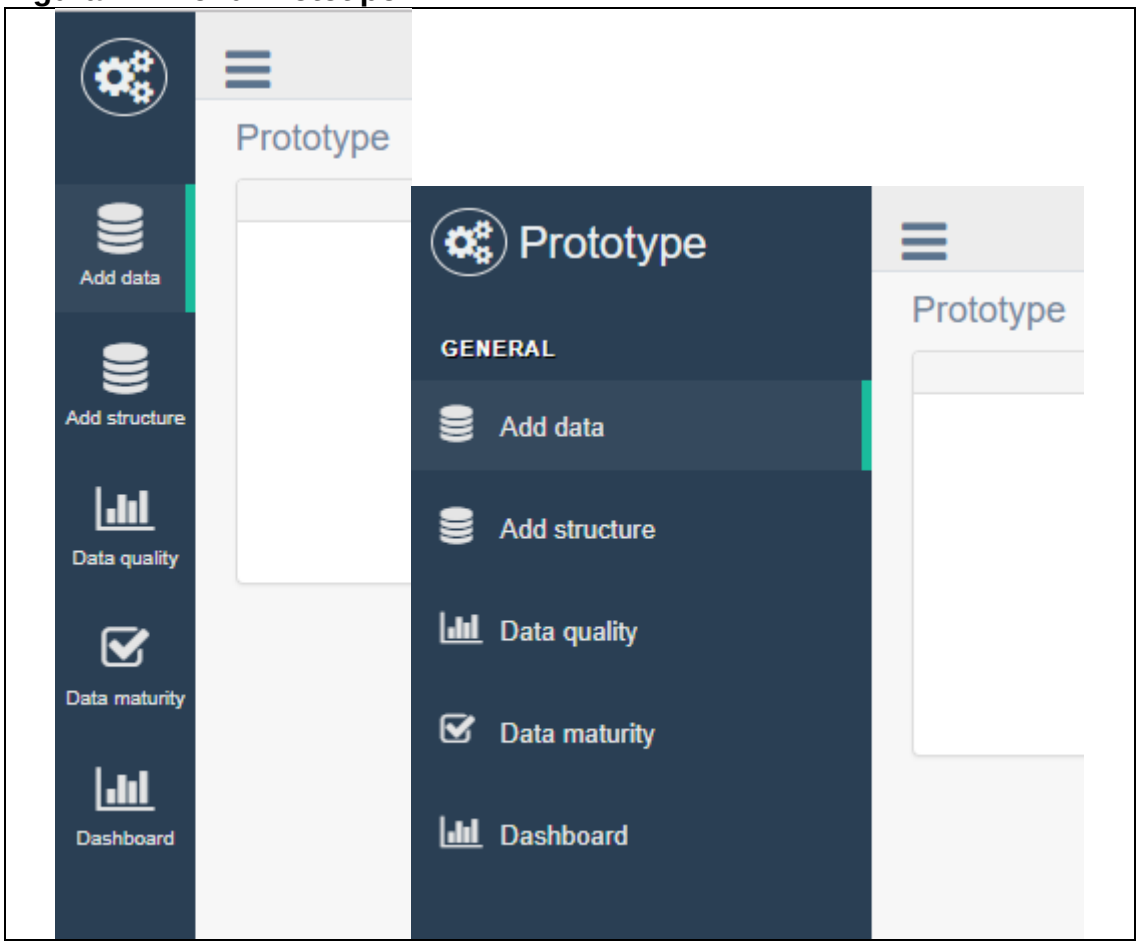
➢El sistema debe proporcionar una interfaz de fácil acceso para el usuario. (Usabilidad)

➢Se espera que el sistema tenga un tiempo de respuesta aceptable debido a la cantidad de datos que debe manejar.

### 11.7 MOCKUPS

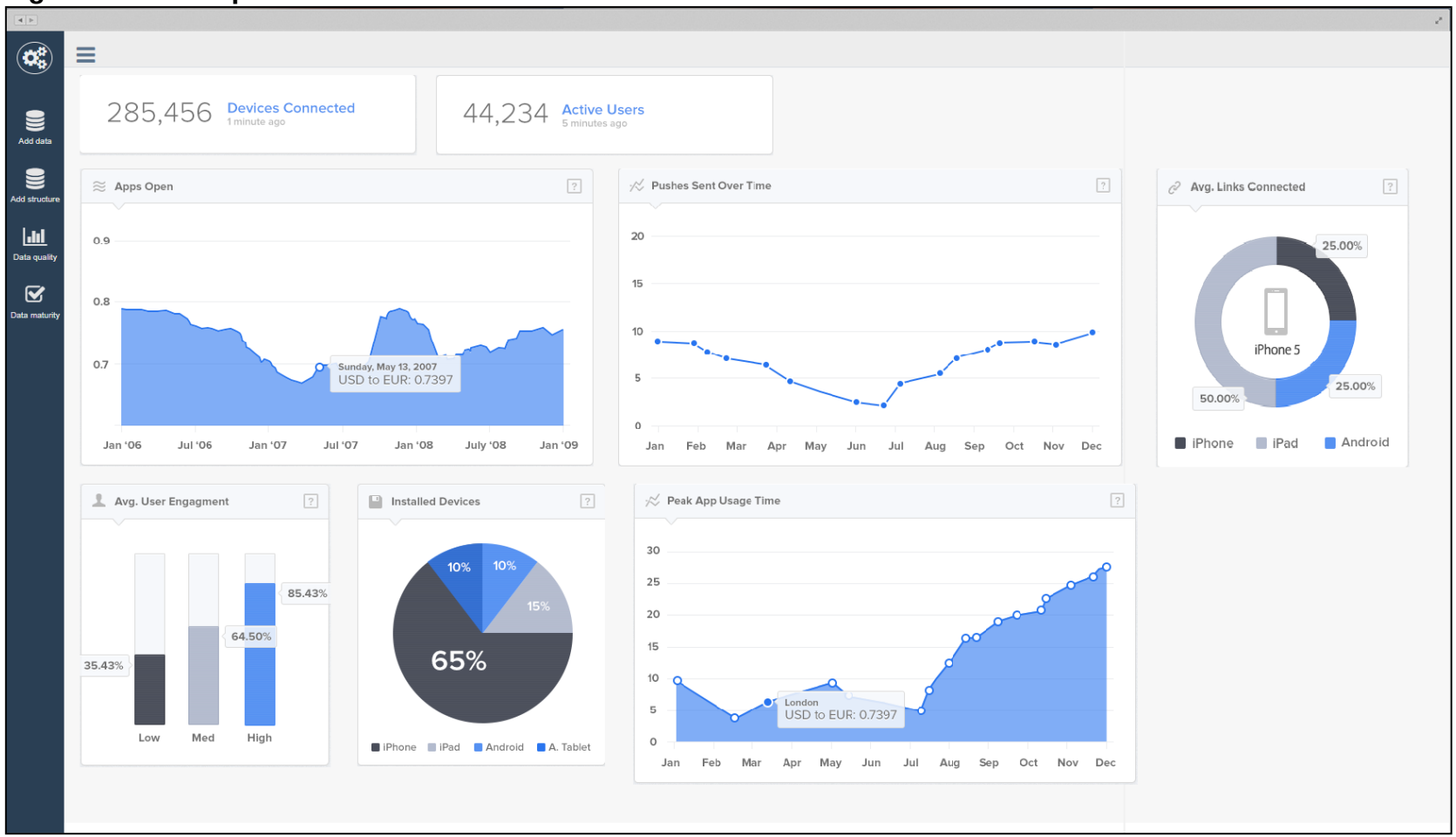
A las opciones del menú lateral de la aplicación existentes para realizar la extracción de la estructura, extracción de datos de una fuente y análisis de los datos. Se incluirá la opción dashboard que redireccionará a la visualización de esta opción, como se visualiza en la Figura 21.

**Figura 21. Menú Prototipo**




Fuente. El Autor

**Figura 22. Mockup Dashboard**




Fuente. El Autor

En el dashboard se incluirán diferentes tipos de gráficos que describirán los datos como:

- Gráfico de barras
- Gráfico de líneas
- Grafico circular
- Mapa
- Gráfico de burbuja

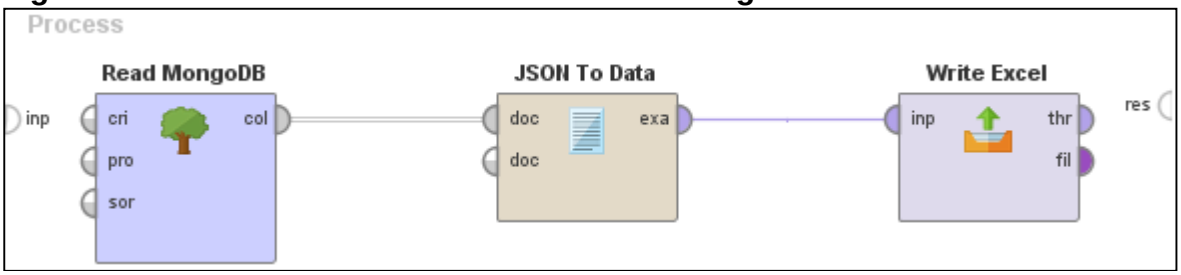
Estos tipos de graficas son embebidas a la aplicación desde Tableau Public donde son alojadas.

El dashboard permite que los datos sean el elemento central de la toma de decisiones al usarlos para contar una historia, impactando la comprensión comercial y respuestas a preguntas en un menor tiempo.

## 12. APLICACIÓN DEL MODELO

Para los valores categóricos se requiere realizar un proceso que consiste en dar un número a cada valor correspondiente utilizándolo como un identificador. Seguido de tener todos los valores numéricos se realiza la normalización de estos. (Cherevko & Malikov, n.d.) Antes de la aplicación de DBSCAN, se debe encontrar el valor épsilon optimo que se encuentra por medio de una gráfica de k-distancias (véase la Figura 23).

**Figura 23. Proceso de Transformación de los Registros de Formato**




Fuente. El Autor

En la Figura 23 se muestra en el proceso realizado para cambiar el formato que consiste en extraer los datos desde mongoDB teniendo como resultado un JSON, se realiza el proceso de transformar de JSON a formato tabular y como último paso se guarda en un archivo Excel para su posterior uso. Este paso se realiza porque los valores en la base datos están todos en formato cadena provocando que en Rapidminer los trate todos de este tipo y al aplicar algunos procesos en la aplicación no se obtiene el resultado esperado.

### 12.1 CONJUNTO DE DATOS SECOP I

Para la aplicación del modelo los campos disponibles del conjunto de datos SECOP I se listan a continuación (véase el Cuadro 4).

**Cuadro 4. Columnas Conjunto de Datos SECOP I**

| # | Columna |
|---|---|
| 1 | fecha_de_firma_del_contrato |
| 2 | fecha_fin_ejec_contrato |
| 3 | fecha_ini_ejec_contrato |
| 4 | id_clase |
| 5 | id_familia |
| 6 | id_grupo |
| 7 | id_objeto_a_contratar |
| 8 | id_origen_de_los_recursos |
| 9 | id_regimen_de_contratacion |
| 10 | id_sub_unidad_ejecutora |
| 11 | id_tipo_de_proceso |

**Cuadro 4. (Continuación)**

| # | Columna |
|---|---|
| 12 | razón_nombre_contratista |
| 13 | municipio_entrega |
| 14 | municipio_obtencion |
| 15 | municipios_ejecucion |
| 16 | nit_de_la_entidad |
| 17 | nivel_entidad |
| 18 | nombre_sub_unidad_ejecutora |
| 19 | orden_entidad |
| 20 | origen_de_los_recursos |
| 21 | plazo_de_ejec_del_contrato |
| 22 | rango_de_ejec_del_contrato |
| 23 | regimen_de_contratacion |
| 24 | tiempo_adiciones_en_dias |
| 25 | tipo_de_contrato |
| 26 | tipo_de_proceso |
| 27 | valor_contrato_con_adiciones |
| 28 | valor_total_de_adiciones |

Fuente. El Autor

Del listado de campos se deben escoger los más relevantes con el propósito de encontrar relaciones entre ellos, Se seleccionaron los siguientes del listado (véase el Cuadro 5):

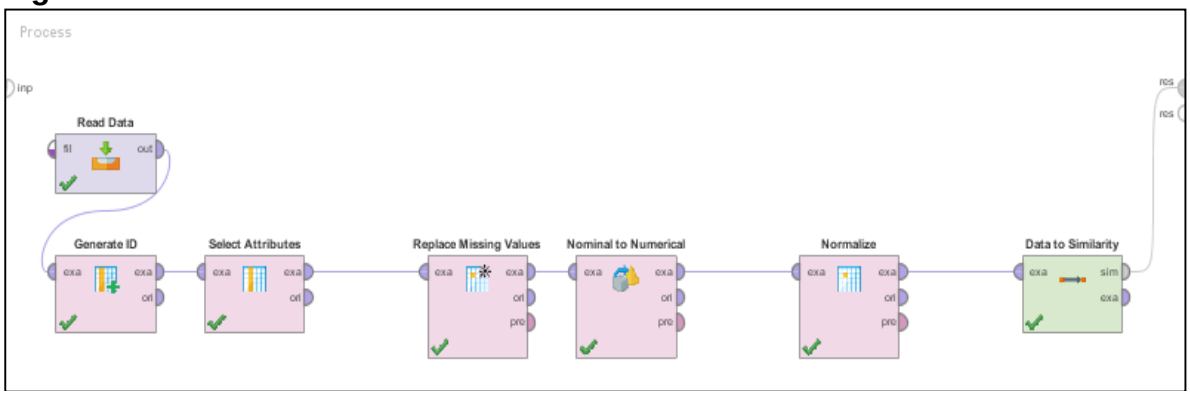
**Cuadro 5. Columnas Seleccionadas del Conjunto de Datos SECOP I**

| # | Columna |
|---|---|
| 1 | departamento_ejecucion |
| 2 | dias_ejec_contrato |
| 3 | nombre_familia |
| 4 | nombre_grupo |
| 5 | objeto_a_contratar |
| 6 | orden_entidad |
| 7 | plazo_de_ejec_del_contrato |
| 8 | tipo_de_contrato |
| 9 | tipo_de_proceso |

Fuente. El Autor

Se realiza el cálculo de épsilon mediante un operador de Rapidminer (véase la Figura 24):

**Figura 24. Proceso de Cálculos de las K-distancias**

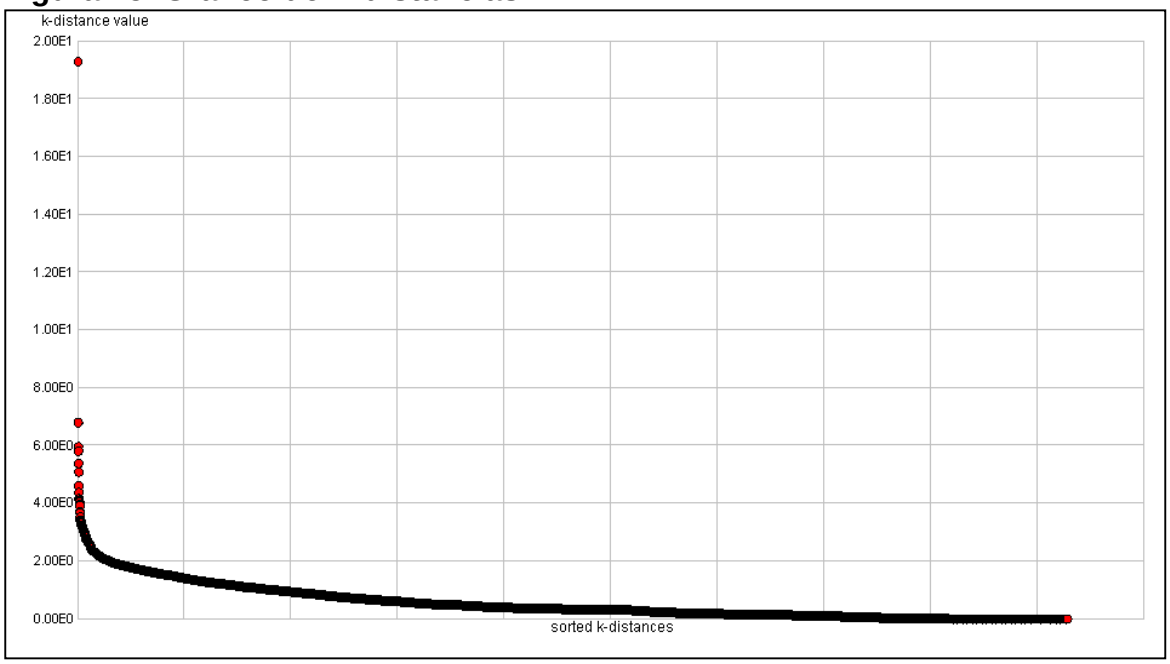



Fuente. El Autor

En la Figura 24se observa la selección de los atributos, reemplazo de valores faltantes, transformación de los valores categóricos a numéricos; posteriormente se normalizan y pasan al proceso Data Similarity que es el encargado de medir las distancias entre cada punto.

En el resultado del anterior proceso observamos en la Figura 225 la gráfica resultante con un k=5.

**Figura 25. Gráfico de K-distancias**

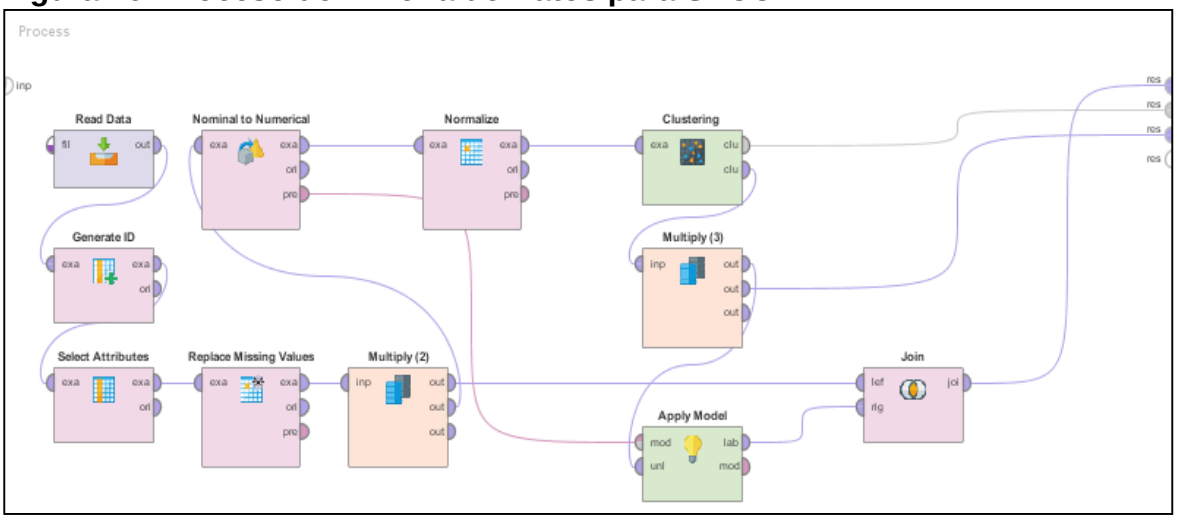



Fuente. El Autor

Épsilon que da como resulta es Epsilon=1.03 y MinPoints= 5.

La aplicación del proceso de agrupamiento DBSCAN en Rapidminer es definido a continuación (véase la Figura 26).

**Figura 26. Proceso de Minería de Datos para SECOP II**




Fuente. El Autor

En la Figura 26 se observan los siguientes pasos:

➢Lectura de los datos (Read Data).

➢Generación de un Id a cada fila (Generate ID).

➢Selección de los atributos con lo que se va a realizar el proceso de minería de datos (Select Attributes).

➢Reemplazo de los valores faltantes debido a que el proceso no soporta valores nulos (Replace Missing Values).

➢Proceso de transformación de valores Nominales a numéricos (Nomina to Numerical).

➢Normalización de los valores numéricos (Normalize).

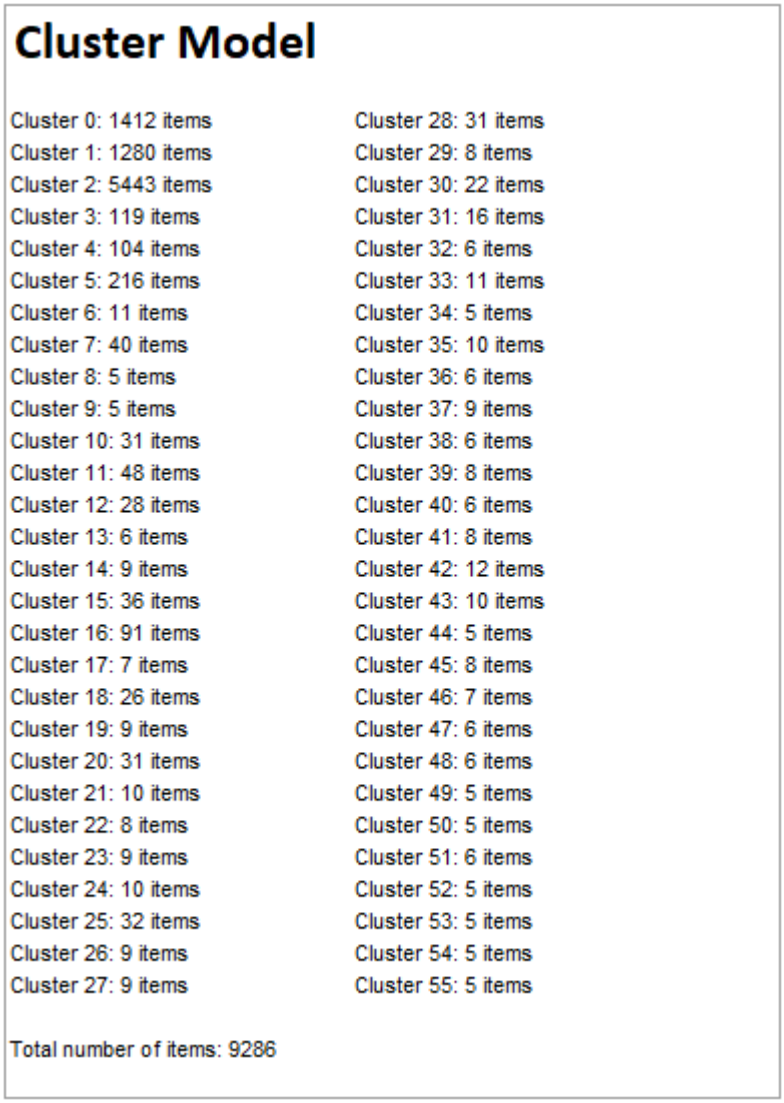
➢Aplicación del proceso de minería de datos DBSCAN (Clustering).

➢Des normalización de valores (Apply Model).

➢Cruce de las filas del resultado con las iniciales para obtener los valores Nominales (Join).

Como resultado da el siguiente agrupamiento (véase la Figura 27).

**Figura 27. Resultado Proceso de Agrupamiento**



Cluster Model

| | |
|---|---|
| Cluster 0: 1412 items | Cluster 28: 31 items |
| Cluster 1: 1280 items | Cluster 29: 8 items |
| Cluster 2: 5443 items | Cluster 30: 22 items |
| Cluster 3: 119 items | Cluster 31: 16 items |
| Cluster 4: 104 items | Cluster 32: 6 items |
| Cluster 5: 216 items | Cluster 33: 11 items |
| Cluster 6: 11 items | Cluster 34: 5 items |
| Cluster 7: 40 items | Cluster 35: 10 items |
| Cluster 8: 5 items | Cluster 36: 6 items |
| Cluster 9: 5 items | Cluster 37: 9 items |
| Cluster 10: 31 items | Cluster 38: 6 items |
| Cluster 11: 48 items | Cluster 39: 8 items |
| Cluster 12: 28 items | Cluster 40: 6 items |
| Cluster 13: 6 items | Cluster 41: 8 items |
| Cluster 14: 9 items | Cluster 42: 12 items |
| Cluster 15: 36 items | Cluster 43: 10 items |
| Cluster 16: 91 items | Cluster 44: 5 items |
| Cluster 17: 7 items | Cluster 45: 8 items |
| Cluster 18: 26 items | Cluster 46: 7 items |
| Cluster 19: 9 items | Cluster 47: 6 items |
| Cluster 20: 31 items | Cluster 48: 6 items |
| Cluster 21: 10 items | Cluster 49: 5 items |
| Cluster 22: 8 items | Cluster 50: 5 items |
| Cluster 23: 9 items | Cluster 51: 6 items |
| Cluster 24: 10 items | Cluster 52: 5 items |
| Cluster 25: 32 items | Cluster 53: 5 items |
| Cluster 26: 9 items | Cluster 54: 5 items |
| Cluster 27: 9 items | Cluster 55: 5 items |

Total number of items: 9286

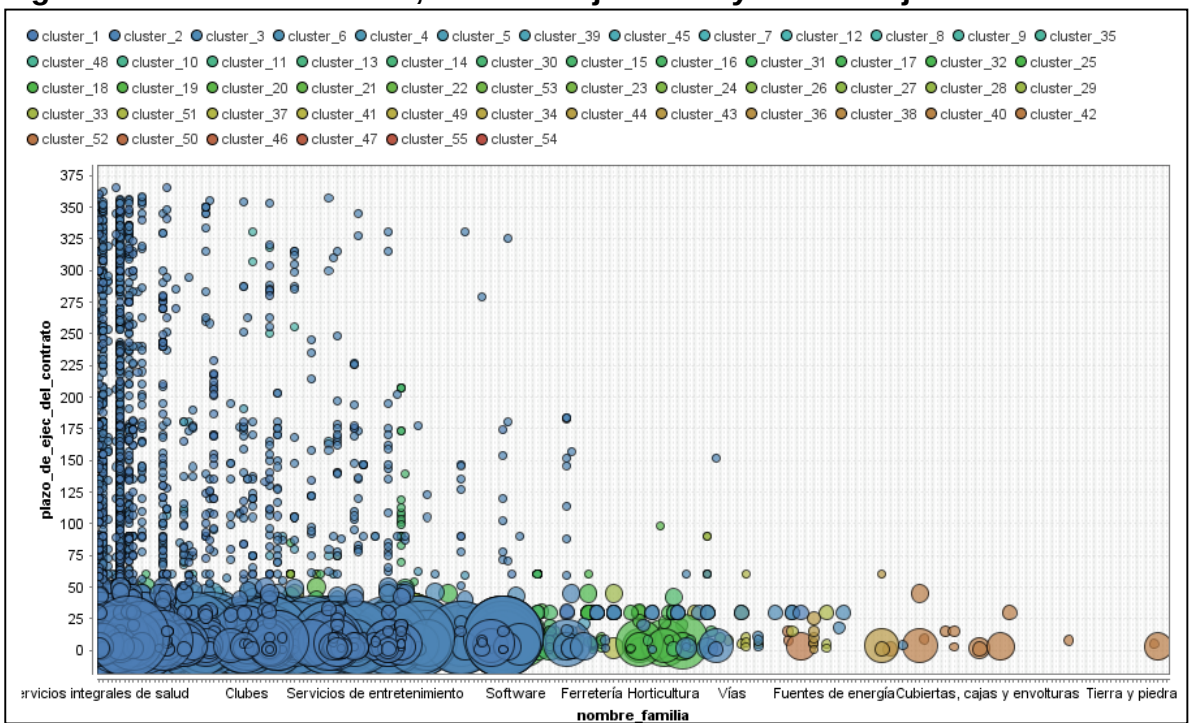
Fuente. El Autor

Como se observa en la Figura 27 se encuentran 55 agrupaciones, un número muy alto respecto a lo recomendado de un máximo de 10 agrupaciones. Un 15% de los elementos se consideran ruido, los grupos 1 y 2 concentran un 72%, el restante se reparte entre los demás grupos siendo aproximadamente un 12%.

## 13. ANÁLISIS DE LOS RESULTADOS

Una vez realizado el proceso de agrupamiento utilizando DBSCAN se obtiene como resultado lo siguiente (véase la Figura 28).

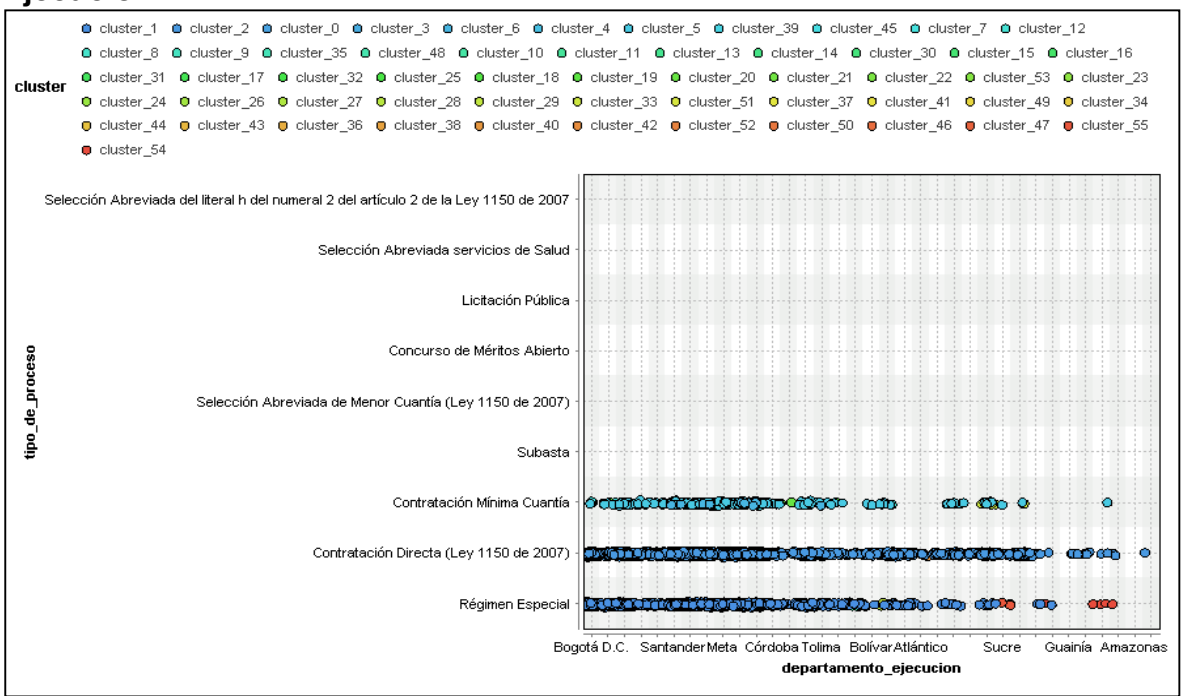
**Figura 28. Relación Familia, Plazo de Ejecución y Días de Ejecución Real**




Fuente. El Autor

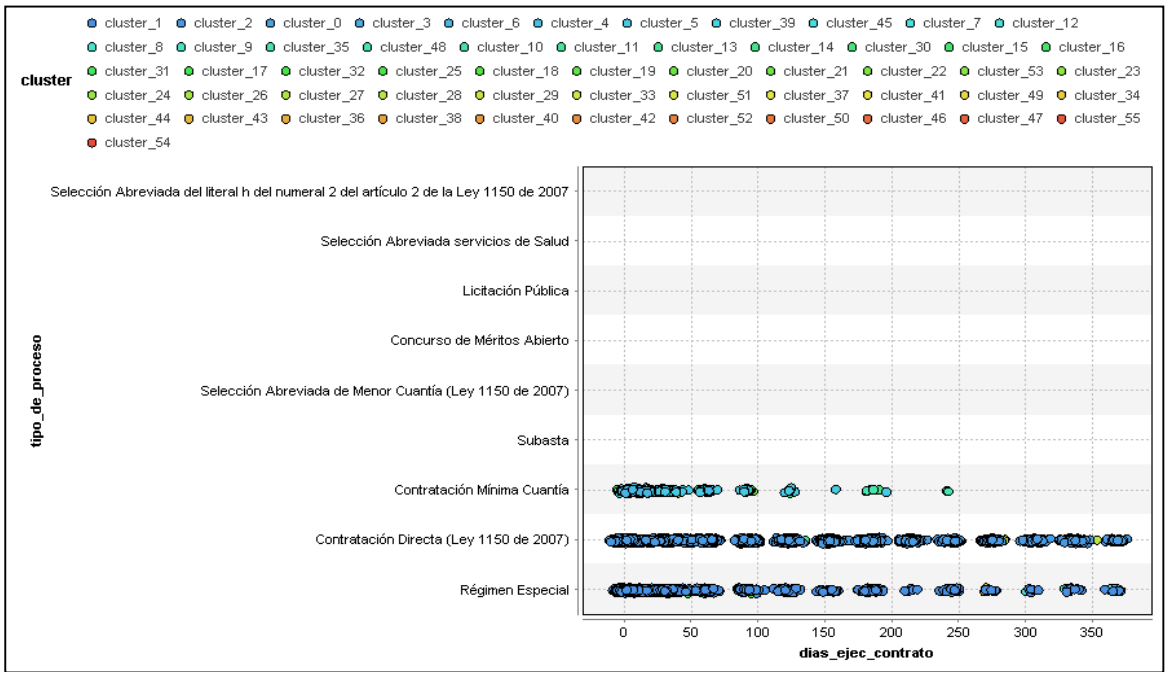
En la Figura 28 se observa que los contratos que tienen un mayor número de días de ejecución real tienen menores tiempo de plazo de ejecución del contrato pertenecientes a las familias de Servicios integrales de salud, Clubes, Servicios de entretenimiento y Software. Significando demoras en el cumplimiento que puede darse por varias causas como aplazamiento de los contratos y demora real en la ejecución del contrato.

**Figura 29. Análisis en el Contexto del Tipo de Proceso – Departamento de Ejecución**



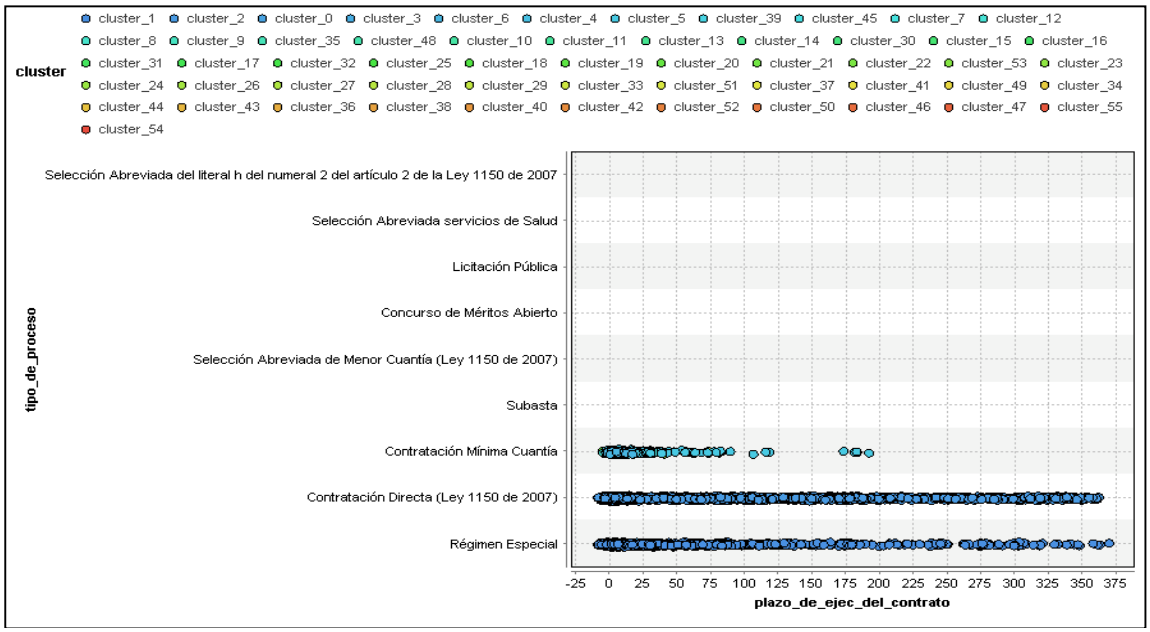

Fuente. El Autor

**Figura 30. Análisis en el Contexto del Tipo de Proceso – Días de Ejecución del Contrato**




Fuente. El Autor

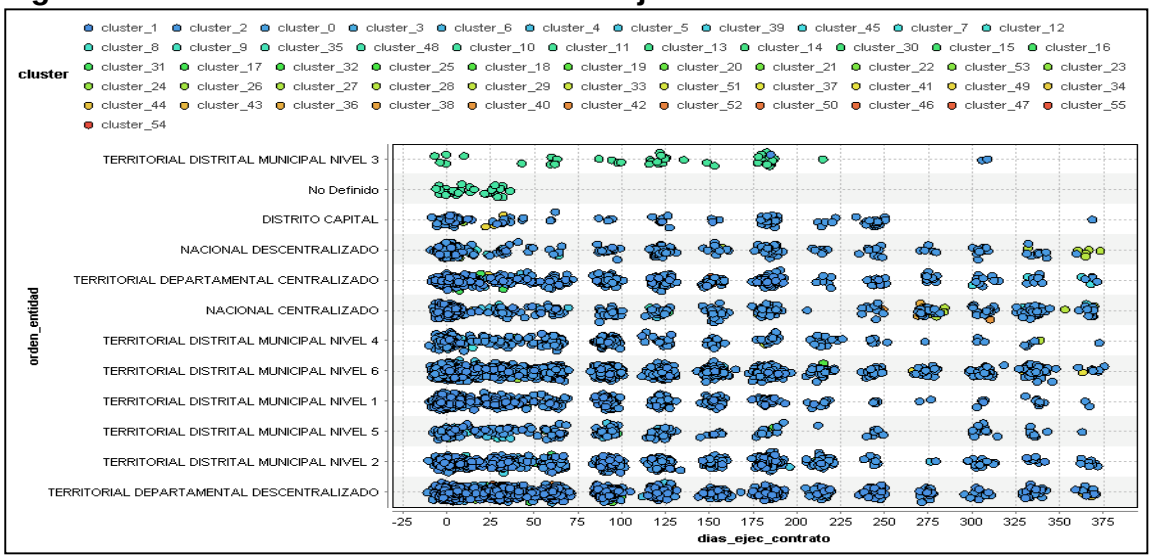
**Figura 31. Análisis en el Contexto del Tipo de Proceso – Plazo de Ejecución del Contrato**




Fuente. El Autor

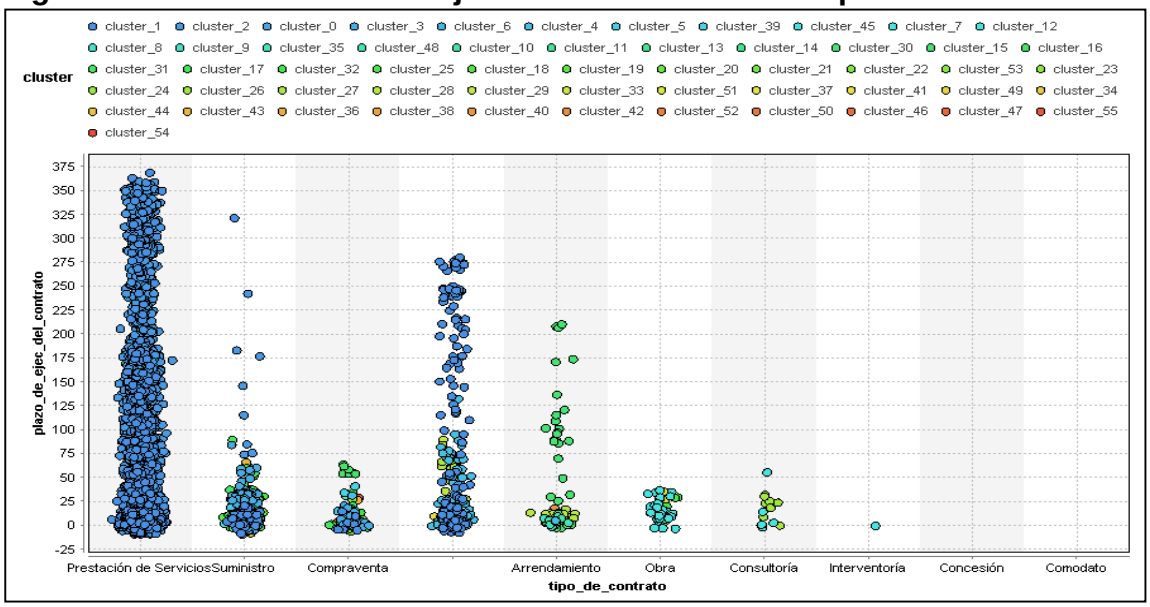
Realizando un análisis entre la Figura 29, Figura 30 y Figura 31 la contratación de mínima cuantía se encuentra agrupado en un número de días menor a 100 días, principalmente usada en Bogotá D.C., Santander, Meta, Córdoba y Tolima. Contrastándolo con el plazo definido para la ejecución del contrato se obtiene una alta relación con el cumplimiento en estas zonas, cabe resaltar que algunos puntos en este grupo no cumplen a cabalidad con esta relación.

**Figura 32. Orden de la Entidad - Días de Ejecución del Contrato**




Fuente. El Autor

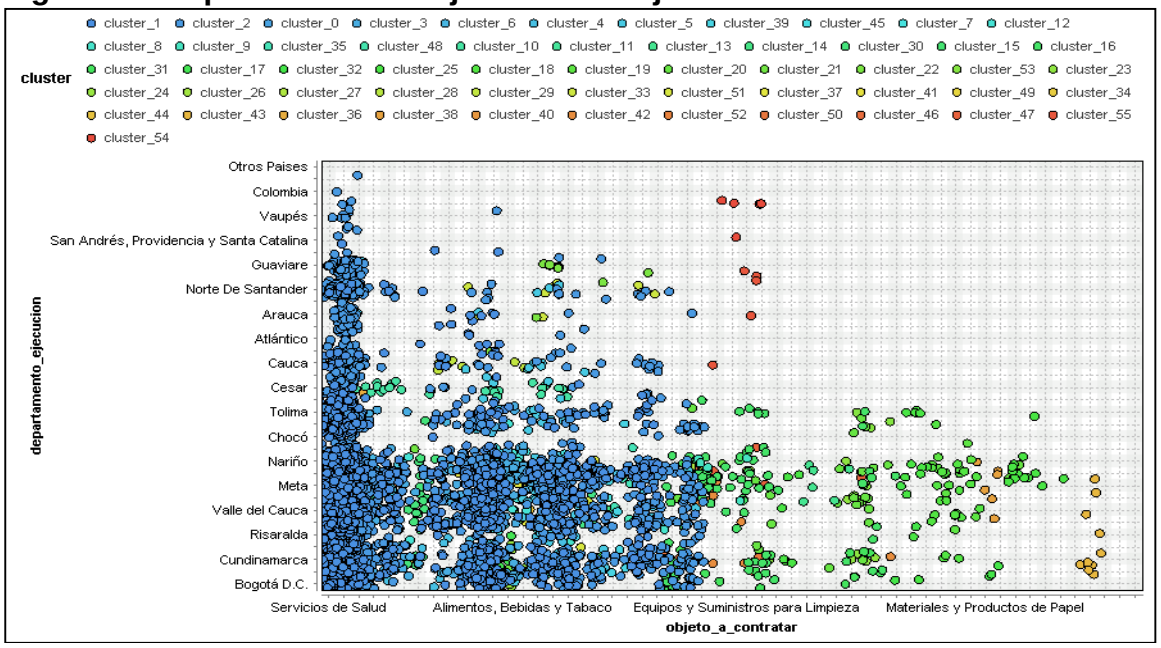
**Figura 33. Análisis Plazo de Ejecución del Contrato - Tipo de Contrato**




Fuente. El Autor

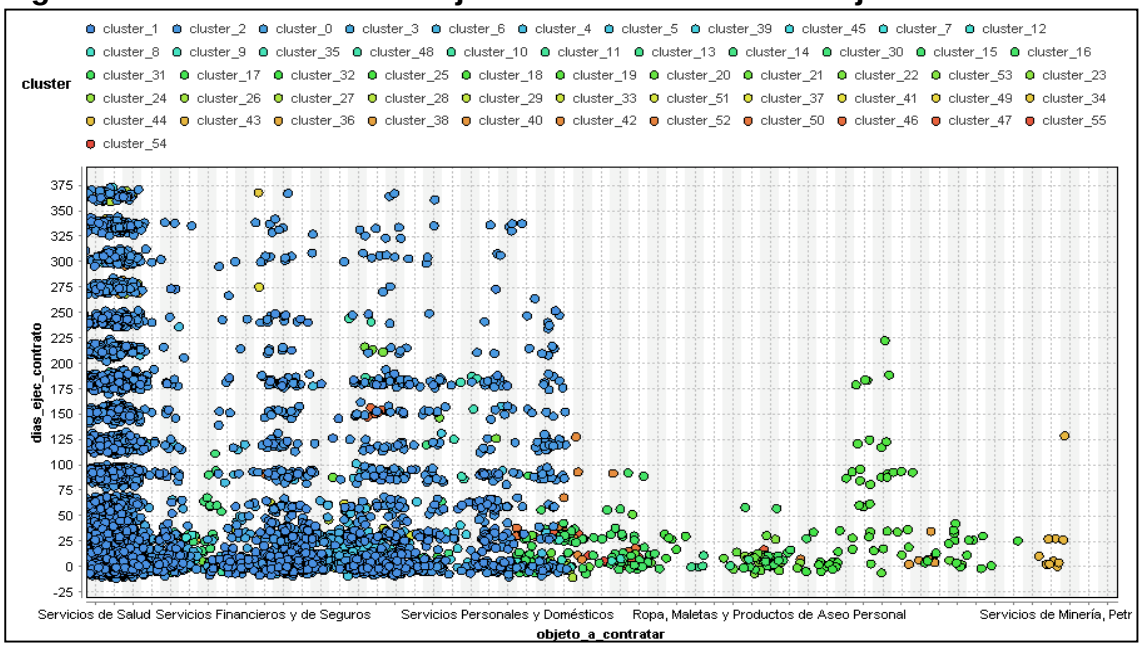
Realizando un análisis entre la Figura 32 y la Figura 33 se observa que las entidades de orden Distrital Municipal de Nivel 3 con una ejecución que comprende entre 1 día a 175 días son relacionados a contratos de arrendamiento principalmente con plazo de ejecución similar a descrito en la ejecución real implicando una tasa aceptable de cumplimiento para este tipo de entidades.

**Figura 34. Departamento de ejecución - Objeto a contratar**



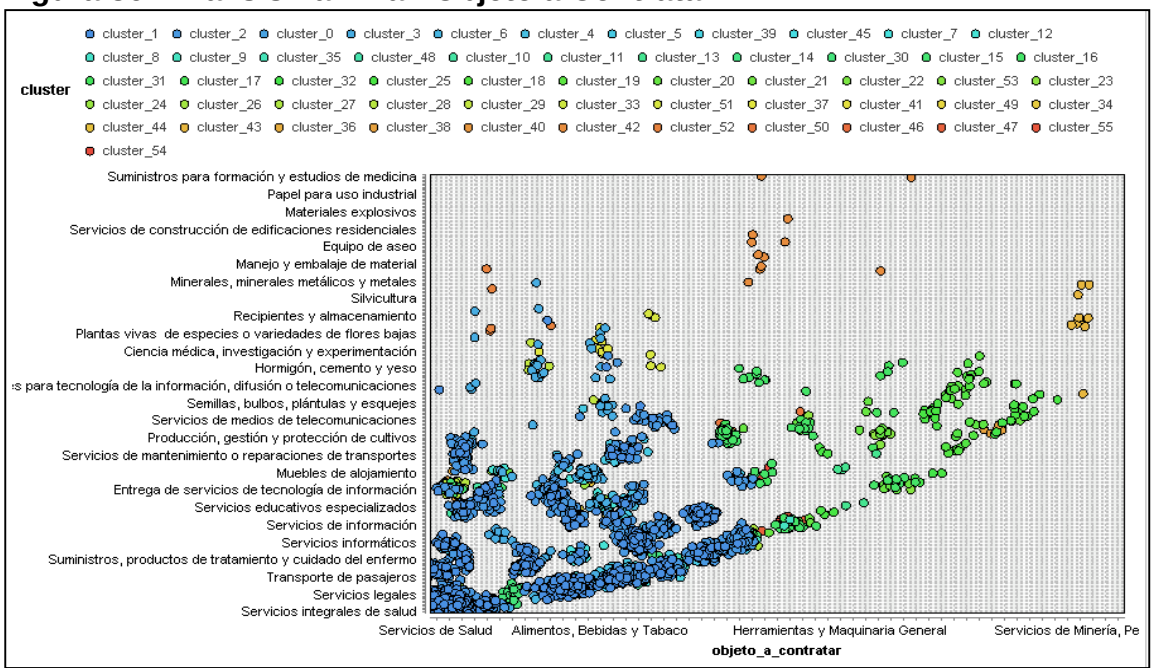

Fuente. El Autor

**Figura 135. Análisis Días de Ejecución del Contrato - Objeto a Contratar**




Fuente. El Autor

**Figura 36. Análisis Familia - Objeto a Contratar**




Fuente. El Autor

Realizando un análisis en la Figura 34, Figura 35 y Figura 36 los clústeres de color verde se encuentra un número de días de ejecución real relativamente bajo entre 1 y 125 días pertenecientes principalmente a Ropa, Maletas y productos de aseo

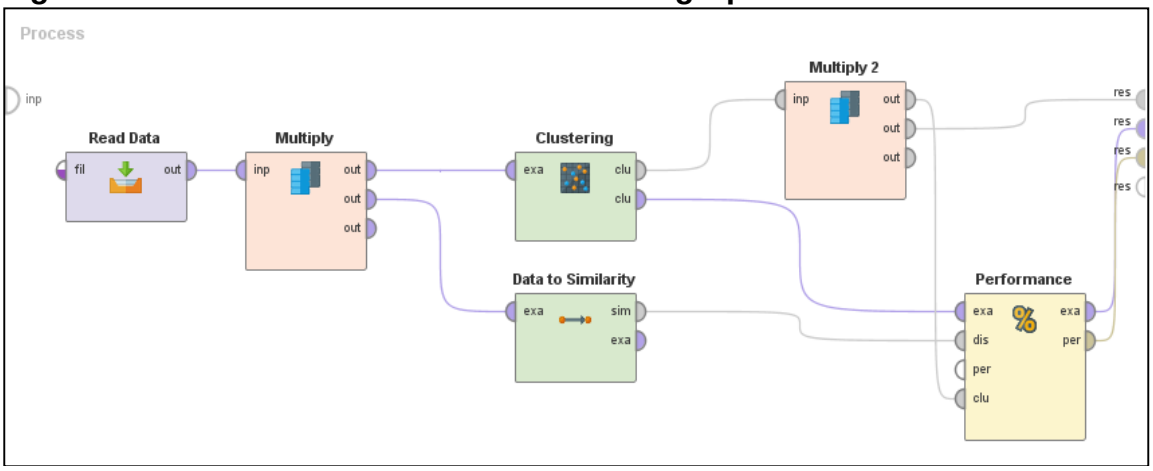
personal y a Herramientas y Maquinaria General en los departamentos de Tolima, Meta, Valle de Cauca y Cundinamarca. Entendiéndose que en estas zonas se ejecutan un alto número de contratos cortos de hasta un máximo de 4 meses en diferentes áreas como Entrega de servicios de tecnología de telecomunicación y en su mayoría son de tipo arrendamiento.

## 14. EVALUACIÓN DEL MODELO

En la clasificación supervisada, la evaluación del modelo de clasificación resultante es una parte integral del proceso de desarrollo de un modelo de clasificación, y existen medidas y procedimientos de evaluación bien aceptados, por ejemplo, precisión y validación cruzada, respectivamente. Sin embargo, debido a su naturaleza, la evaluación de agrupaciones no es una parte bien desarrollada o comúnmente utilizada del análisis de agrupaciones. Sin embargo, la evaluación de conglomerados o la validación de conglomerados es de gran importancia.

La evaluación del modelo se va a realizar por medio de un operador que provee Rapidminer llamado Performance, permitiendo medir el promedio de distancias de los elementos agrupados en cada clúster. A continuación, se muestra el proceso a realizar en el que solo encontramos como diferencia el operador nombrado anteriormente (véase la Figura 37).

**Figura 37. Evaluación del Rendimiento del Agrupamiento**

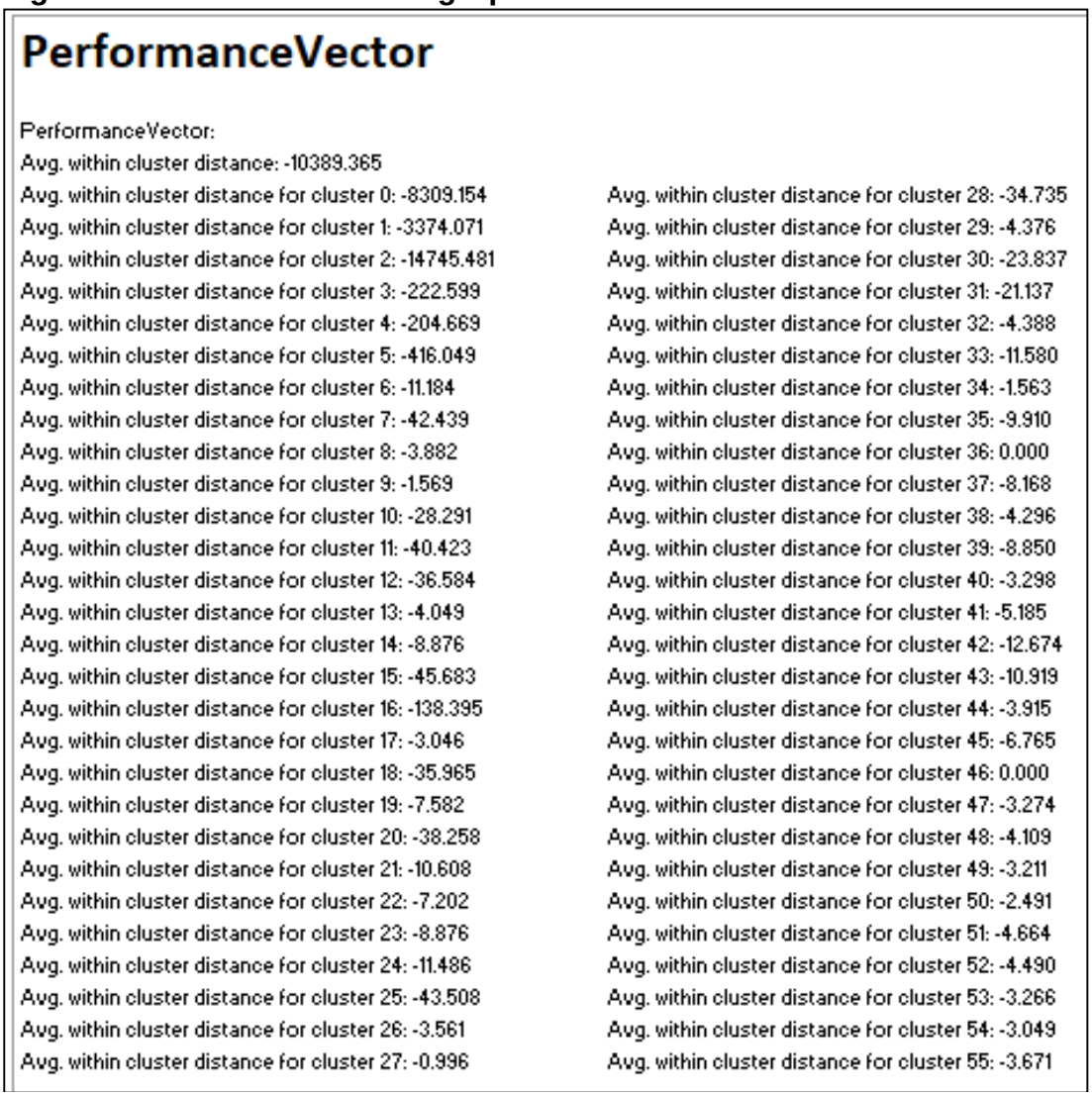



Fuente. El Autor

En la Figura 37 se observa que los datos requeridos para calcular el rendimiento del agrupamiento requieren las distancias entre cada punto utilizado anteriormente para el cálculo de Epsilon y las salidas del Clustering: Información general del agrupamiento y el conjunto de datos con la etiqueta cluster.

El resultado de salida es el siguiente (véase la Figura 38).

**Figura 38. Rendimiento del Agrupamiento**



PerformanceVector

PerformanceVector:
Avg. within cluster distance: -10389.365

| | |
|---|---|
| Avg. within cluster distance for cluster 0: -8309.154 | Avg. within cluster distance for cluster 28: -34.735 |
| Avg. within cluster distance for cluster 1: -3374.071 | Avg. within cluster distance for cluster 29: -4.376 |
| Avg. within cluster distance for cluster 2: -14745.481 | Avg. within cluster distance for cluster 30: -23.837 |
| Avg. within cluster distance for cluster 3: -222.599 | Avg. within cluster distance for cluster 31: -21.137 |
| Avg. within cluster distance for cluster 4: -204.669 | Avg. within cluster distance for cluster 32: -4.388 |
| Avg. within cluster distance for cluster 5: -416.049 | Avg. within cluster distance for cluster 33: -11.580 |
| Avg. within cluster distance for cluster 6: -11.184 | Avg. within cluster distance for cluster 34: -1.563 |
| Avg. within cluster distance for cluster 7: -42.439 | Avg. within cluster distance for cluster 35: -9.910 |
| Avg. within cluster distance for cluster 8: -3.882 | Avg. within cluster distance for cluster 36: 0.000 |
| Avg. within cluster distance for cluster 9: -1.569 | Avg. within cluster distance for cluster 37: -8.168 |
| Avg. within cluster distance for cluster 10: -28.291 | Avg. within cluster distance for cluster 38: -4.296 |
| Avg. within cluster distance for cluster 11: -40.423 | Avg. within cluster distance for cluster 39: -8.850 |
| Avg. within cluster distance for cluster 12: -36.584 | Avg. within cluster distance for cluster 40: -3.298 |
| Avg. within cluster distance for cluster 13: -4.049 | Avg. within cluster distance for cluster 41: -5.185 |
| Avg. within cluster distance for cluster 14: -8.876 | Avg. within cluster distance for cluster 42: -12.674 |
| Avg. within cluster distance for cluster 15: -45.683 | Avg. within cluster distance for cluster 43: -10.919 |
| Avg. within cluster distance for cluster 16: -138.395 | Avg. within cluster distance for cluster 44: -3.915 |
| Avg. within cluster distance for cluster 17: -3.046 | Avg. within cluster distance for cluster 45: -6.765 |
| Avg. within cluster distance for cluster 18: -35.965 | Avg. within cluster distance for cluster 46: 0.000 |
| Avg. within cluster distance for cluster 19: -7.582 | Avg. within cluster distance for cluster 47: -3.274 |
| Avg. within cluster distance for cluster 20: -38.258 | Avg. within cluster distance for cluster 48: -4.109 |
| Avg. within cluster distance for cluster 21: -10.608 | Avg. within cluster distance for cluster 49: -3.211 |
| Avg. within cluster distance for cluster 22: -7.202 | Avg. within cluster distance for cluster 50: -2.491 |
| Avg. within cluster distance for cluster 23: -8.876 | Avg. within cluster distance for cluster 51: -4.664 |
| Avg. within cluster distance for cluster 24: -11.486 | Avg. within cluster distance for cluster 52: -4.490 |
| Avg. within cluster distance for cluster 25: -43.508 | Avg. within cluster distance for cluster 53: -3.266 |
| Avg. within cluster distance for cluster 26: -3.561 | Avg. within cluster distance for cluster 54: -3.049 |
| Avg. within cluster distance for cluster 27: -0.996 | Avg. within cluster distance for cluster 55: -3.671 |

Fuente. El Autor

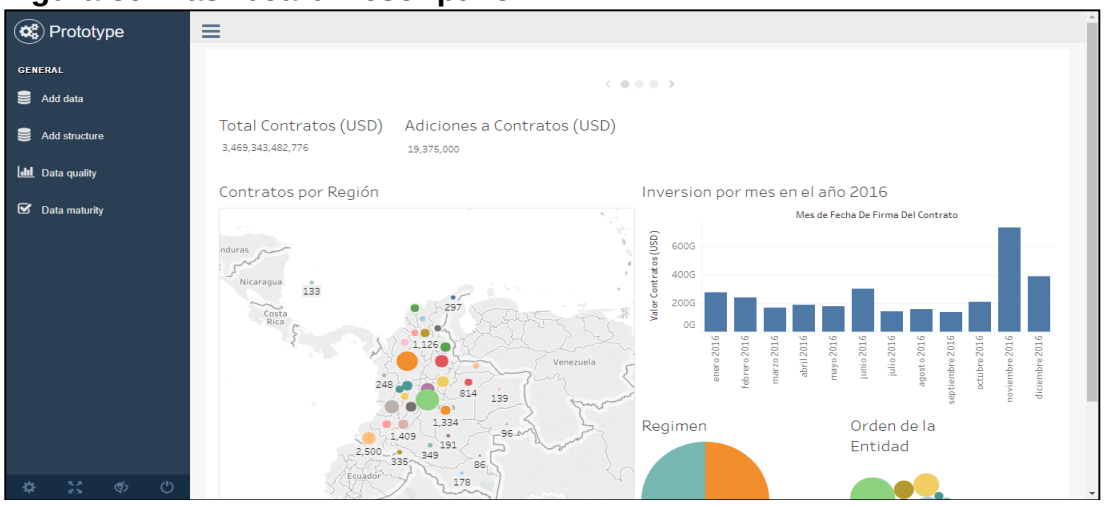
En la Figura 38 se observa que en los agrupamientos 1 y 2 la distancia promedio es alta principalmente en el 2, indicando dispersión entres los puntos de este grupo.

La forma más confiable de evaluar agrupamientos es revisando los datos, si los agrupamientos son útiles y tienen sentido para un experto en el tema.

## 15. DESPLIEGUE

En el Dashboard descriptivo permite realizar varios filtros con la información como lo son: región del contrato, mes de inversión, régimen y orden de la entidad (véase la Figura 39).

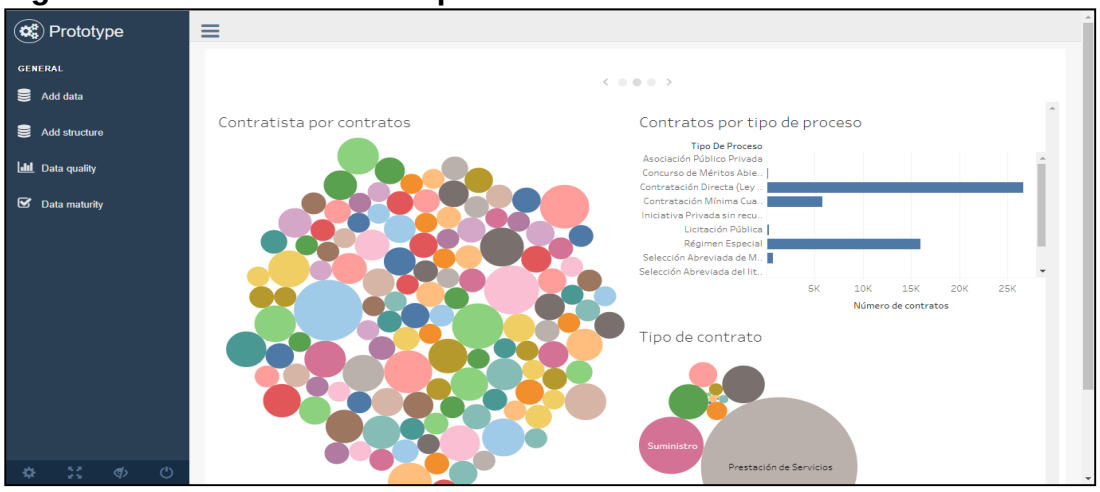
**Figura 39. Dashboard Descriptivo - 1**




Fuente. El Autor

En la siguiente parte del Dashboard tiene los siguientes filtros: Contratista, tipo de proceso, tipo de contrato (véase la Figura 40).

**Figura 40. Dashboard Descriptivo – 2**




Fuente. El Autor

El uso de este tipo de Dashboard permite realizar un primer acercamiento al contexto y a la exploración de los datos.

# 16. CONCLUSIONES

Mediante criterios se seleccionó el método de minería de datos más adecuado para este contexto, con un enfoque en los algoritmos de agrupación que son los que más se ajustan al objetivo de encontrar relaciones nuevas en los datos, se elige el algoritmo DBSCAN debido a su naturaleza de tener tolerancia a datos atípicos que en este caso por medio de la exploración se detectó un alto porcentaje. También se tuvo en cuenta que este algoritmo no funciona adecuadamente con un numero de atributos alto por lo que se optó por seleccionar los 9 atributos más relevantes.

El Diseño de la arquitectura finalizo satisfactoriamente con base en la arquitectura propuesta en la herramienta de extracción de datos cliente-servidor, se incluyó un nuevo componente que se conecta a la interfaz gráfica existente donde el usuario puede realizar una exploración de los datos y visualización del resultado de minería de datos ejecutado al conjunto de datos SECOP I.

El desarrollo del prototipo diseñado queda implementado en un servidor de la Universidad–sede Claustro. El prototipo consiste en agregar un dashboard descriptivo y uno donde se visualizará el resultado del agrupamiento realizado con el algoritmo DBSCAN. Siguiendo la metodología CRISP-DM para la aplicación de la minería de datos se realizó una limpieza, exploración y entendimiento de los datos para la construcción del modelo. Seguido a esto se prosiguió con la ejecución del algoritmo, análisis y evaluación del mismo.

En los resultados de la agrupación se observan algunos grupos que tienen una relación entre días de ejecución que duro y plazo de ejecución del contrato pactado, en donde regiones como Bogotá D.C., Santander, Meta, Córdoba y Tolima es usada en un alto porcentaje la contratación de mínima cuantía en comparación de las demás zonas y se refleja un menor retraso en la ejecución de estos.

La evaluación de la calidad de datos se realizó sobre el modelo propuesto de minería de datos utilizando DBSCAN donde se eligieron los parámetros adecuados que minimizaran la distancia entre puntos de un mismo grupo. En la evaluación se obtuvieron grupos que tienen un gran numero elementos, pero a la vez la distancia promedia es alta en comparación con grupos de menor cantidad de elementos.

Los datos proporcionados por Colombia Compra Eficiente del Sistema Electrónico para la Contratación Pública (Secop) desde el 2016 como base de datos invaluable que incluye información de los procesos de contratación públicos ayudan a evidenciar los problemas de transparencia de los procesos de licitación. En Colombia prevalece la contratación directa a pesar de las diversas modalidades, según lo observado en los datos de cada proceso hay un número reducido de participantes y una variedad de modalidades de contratación que agregan complejidad innecesaria.

# 17. RECOMENDACIONES Y TRABAJOS FUTUROS

Para la continuación del proyecto se recomienda realizar una revisión minuciosa de los datos extraídos de datos abierto, debido a su naturaleza traen columnas sin estandarizar que impactan el análisis de los mismos. El mejoramiento en la extracción de valor en los datos y el análisis hacen parte fundamental del trabajo y que debido a una exploración de los datos sin la profundidad necesaria afecto el resultado de estos. Mediante el uso de la metodología CRISP-DM se completó cada uno de los objetivos planteado, aun así, queda mucho trabajo por hacer en la aplicación de métodos de minería a los datos.

En el tema de las herramientas necesarias para llevar a cabo el proceso de minería y publicación de los resultados se deben analizar herramientas que tengan opciones sin licenciamientos debido a que puede afectar tanto el resultado como el acceso a estos.

En la selección del algoritmo adecuado se debe profundizar en algoritmos especializados en variables categóricas a causa de que los datos suministrados de SECOP en su mayoría son de este tipo y considerarse la elección de al menos 2 de estos para realizar una comparación entre resultados.

## BIBLIOGRAFÍA


CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ, Thomas; SHEARER, Colin y WIRTH, Rüdiger. CRISP-DM 1.0. Step-by-step data mining guide. Pittsburgh: The CRISP-DM consortium, 2000. 76 p.

EDDIB, A.J.A.; MOHAMMED, E.M. y CHAHHOU, M. Algorithms and systems for data mining: A survey. En: Colloquium in Information Science and Technology, CIST. Enero, 2015. no. 2.

FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). From Data Mining to Knowledge Discovery in Databases. En: AI Magazine. Junio – julio, 1996. vol. 17, no. 3.

FU, Tak-chung. A review on time series data mining.en: Engineering Applications of Artificial Intelligence. February, 2011. vol. 24, no. 1.

GOMEZ ZOTANO, Miguel Angel y BERSINI, Hugues. A Data-driven Approach to Assess the Potential of Smart Cities: The Case of Open Data for Brussels Capital Region. En: Energy Procedia. Marzo – abril, 2017. vol. 111.

GORUNESCU, Florin. Data Mining: Concepts and Techniques. Berlin: Morgan Kaufmann, 2011. 360 p.

JIMÉNEZ GALINDO, Álvaro y ÁLVAREZ GARCÍA, Hugo. Minería de Datos en la Educación [en línea]. Madrid: Revista Inteligencia En Redes de Comunicación [citado 15 agosto, 2017]. Disponible en Internet: <URL: https://www.it.uc3m.es /jvillena/irc/practicas/10-11/08mem.pdf>

KITCHIN, R.; LAURIAULT, T.P. y MCARDLE, G. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. En: Regional Studies, Regional Science. Junio – agosto, 2015. vol. 2, no. 1.

KOTU, Vijay y DESHPANDE, Bala. Predictive analytics and data mining: concepts and practice with RapidMiner. Massachusetts, United States of America: Elliot Steven, 2015. ISBN: 978-0-12-801460-8.

LAND, Sebastián y FISCHER, S. RapidMiner in academic use, V, 1-3 [en línea]. Berlín: Rapid-I GmbH [citado 20 agosto, 2017]. Disponible en Internet: <URL: http://docs.rapidminer.com/downloads/RapidMiner _RapidMinerInAcademicUse_en.pdf>

MÁCHOVÁ, R., y LNENICKA, M. Evaluating the Quality of Open Data Portals on the National Level. En: Journal of Theoretical and Applied Electronic Commerce Research. Mayo – junio, 2017. vol. 12, no. 1.

MARISCAL, Gonzalo; MARBÁN, Óscar y FERNÁNDEZ, Covadonga. A survey of data mining and knowledge discovery process models and methodologies. En: The Knowledge Engineering Review. Junio – agosto, 2010. vol. 25, no. 2.

MARTIN, Erika G y BEGANY, Grace M. Opening government health data to the public: benefits, challenges, and lessons learned from early innovators. En: Journal of the American Medical Informatics Association. Agosto – septiembre, 2016. vol. 24, no. 2.

RIQUELME, J.C.; RUIZ, R. y GILBERT, K. Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial, En: Revista Iberoamericana de Inteligencia Artificial. Febrero – marzo, 2006. no, 29.

RUDY; MIRANDA, E., y SURYANI, E. Implementation of datawarehouse, datamining and dashboard for higher education. En: Journal of Theoretical and Applied Information Technology. Junio – julio, 2014. vol. 64, no. 3.

SHAH, Chintan; y JIVANI, Anjali. Comparison of data mining clustering algorithms [en línea]. Nirma: University International Conference on Engineering (NUiCONE) [citado 20 agosto, 2017]. Disponible en Internet: <URL: https://doi.org/10.1109 /NUiCONE.2013.6780101>

SUAKANTO, S.; SUPANGKAT, S.H.; SUHARDI, A. y SARAGIH, R. Smart city dashboard for integrating various data of sensor networks [en línea]. Tangerang: International Conference on ICT for Smart Society [ciado 20 agosto, 2017]. Disponible en Internet: <URL: https://doi.org/10.1109/ICTSS.2013.6588063>

VALENCIA ZAPATA, Gustavo Adolfo. Minería de datos la minería de datos como herramienta para la toma de decisiones estratégicas. Bogotá: IMG, 2010. 18 p.